Array 14 (2022) 100168

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Generalised network architectures for environmental sensing: Case studies for a digitally enabled environment

M.I. Mead [a,b,**], M. Bevilacqua [c], C. Loiseaux [c], S.H. Hallett [b], S. Jude [b], C. Emmanouilidis [d], J. Harris [b], P. Leinster [b], S. Mutnuri [e], T.H. Tran [c,*], L. Williams [c,***]

[a] School of Public Health, Environmental Research Group, Imperial College London, London, SW7 2AZ, UK
[b] Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, MK43 0AL, UK
[c] Centre for Competitive Creative Design, Cranfield University, Cranfield, MK43 0AL, UK
[d] Through-life Engineering Services Institute, Cranfield University, Cranfield, MK43 0AL, UK
[e] Department of Biological Sciences, Birla Institute of Technology and Science, Goa, 403726, India

ARTICLE INFO

Keywords:
Ubiquitous sensor networks
Network analytics
Integrated sensing
Network policy considerations
Living laboratory
Digital environment
Internet of things
Urban observatory

ABSTRACT

A digitally enabled environment is a setting which incorporates sensors coupled with reporting and analytics tools for understanding, observing or managing that environment. Large scale data collection and analysis are a part of the emerging digitally enabled approach for the characterisation and understanding of our environment. It is recognised as offering an effective methodology for addressing a range of complex and interrelated social, economic and environmental concerns. The development and construction of the approach requires advances in analytics control linked with a clear definition of the issues pertaining to the interaction between elements of these systems. This paper presents an analysis of selected issues in the field of analytics control. It also discusses areas of progress, and areas in need of further investigation as sensing networks evolve. Three case studies are described to illustrate these points. The first is a physical analytics test kit developed as a part of the "Reinvent the Toilet Challenge" (RTTC) for process control in a range of environments. The second case study is the Cranfield Urban Observatory that builds on elements of the RTTC and is designed to allow users to develop user interfaces to monitor, characterise and compare a variety of environmental and infrastructure systems plus behaviours (e.g., water distribution, power grids). The third is the Data and Analytics Facility for National Infrastructure, a cloud-based high-performance computing cluster, developed to receive, store and present such data to advanced analytical and visualisation tools.

1. Introduction

Large scale data collection and analysis are increasingly seen as the best means to address issues such as urban planning, climate resilience, digital security, declining air quality and infrastructure development [1,2]. They are also key to the development of the digitally enabled environment (DEE). A DEE can be defined as an integrated system comprising of devices or agents which can interact with each other via a range of communication tools, with associated analytical and data management capabilities. These devices and agents can also be used both for monitoring and management of the DEE itself. The environmental DEE represents the application of this approach in understanding and managing the atmosphere and natural environment.

A broad range of low-cost environmental and latent physical parameter sensing technologies are now widely available. The breadth, variation and relatively low cost of many of these sensors is in part due to the rapid uptake of accessible electronics/computing environments and related developments within Internet of Things (IoT) infrastructures. Observable parameters can include physical, spectroscopic, chemical, biological, biometric, acoustic and vibration, electrical, thermodynamic and thermophysical properties (with measurements taken either directly or indirectly). These can be used to provide insights into the wider environment or specific information about infrastructure or networks. Key to this has been the advent of

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail address: t.h.tran@cranfield.ac.uk (T.H. Tran).

https://doi.org/10.1016/j.array.2022.100168
Received 10 August 2020; Received in revised form 11 February 2022; Accepted 4 April 2022
Available online 8 April 2022

widespread low-cost wireless communications hardware and software, primarily associated with global telecommunications. Early examples are the General Packet Radio Service (GPRS) and Global Navigation Satellite System, e.g., Global Positioning Systems. The GPRS is linked to 2nd and 3rd generation mobile telephony networks (2G and 3G). This has been surpassed by post 3G and 4G standards and approaches, e.g., wideband code division multiple access and long-term evolution, and the emergent 5G infrastructure. These developments have allowed large amounts of geo-temporally referenced data to be telemetered globally on previously unachievable scales and with relative ease. Coupled with advances in encryption and fidelity retention, together with allied cloud capabilities for receiving and processing the data arising, this has further opened new opportunities for broad scale information transfer.

In this paper, we discuss a framework of generalised network architectures for environmental sensing to analyse the digital enabled environmental approach in the field of analytics control. We seek to explore end-to-end aspects of network design, some of the wider landscape of network operation, and the data pipeline. The paper discusses selected case studies specifically from an application perspective, as well as drawing on wider aspects of dispersed heterogeneous environmental sensing networks. The concepts described in this paper will be discussed in the context of 'Urban Observatory' and 'Living Laboratory' type experiments as contemporary. These are examples of applied broad scale multi-disciplinary sensing and with consideration of cloud-based platforms able to receive and hold the data arising and support its subsequent analysis and visualisation.

2. Architecture for dispersed heterogeneous environmental sensing networks

2.1. Observational networks overview

Sensor technologies and approaches are increasingly being combined with cutting edge communication technologies to offer integrated systems for collecting data as well as their subsequent transmission and storage. These are the base requirements for observational networks which in addition require further analysis to understand the collected data and then onward dissemination of results to stakeholders. For integrated observational networks, emergent issues with data quality assessment, quality assurance, data assimilation, security and dissemination of outputs need to be addressed. A typical IoT deployment infrastructure can be described as a layered stack. Starting from the physical layer which initiates data generation workflows, through system management and analytics workflow layers and finally to the application layer. The application layer being where data workflows are converted to value-adding services for a range of users. To deliver value, these stack environments implement a range of connectivity standards which are in turn relevant to different connectivity layers. These mostly follow the Open Systems Interconnection layers [3] guidelines but require cross-layer interoperability for effective integration, including the application layer and semantic interoperability. The conversion of data workflows is in most cases associated with applying intelligent analytics to suggest contextually relevant actions, wherein context is determined by the circumstances and the environment that the acquired data are derived from. Coupled with the need for an infrastructure that is easy to procure, install and maintain at low cost, new approaches are needed to model the context of data to deliver situational awareness and to drive the development and operation of transformative sensing approaches.

2.2. Generalised network topology models

The traditional model for analytics control and monitoring is to instrument the observation subject and to develop a data feed directed at a designated control/observation point. Incorporating multiple sensors and various forms of wireless communications represents an extension of this approach.

Four broad high-level network topologies are outlined in Fig. 1 and described below in further detail. Here, the local node refers to controlling electronics which may also include a form of local data or settings storage. The access node can be used for automatic data telemetry, direct local access, or other forms of interrogative access. The sensor/actuator, local node and access node can be in a single enclosure, or as components situated remotely.

2.3. Network data

A common consequence of these emerging sensor networks is the generation of huge amount of data, often referred to as big data. Big data describes information/data assets characterised by a number of 'V's, their high 'volume', 'velocity', 'variety', 'variability', and 'value' for different stakeholders. It requires high-capacity cloud-processing services [5,6], as well as the associated analytical approaches, such as machine learning and wider AI approaches. This involves the ability to store, process and analyse both structured and unstructured data, combining archive and real-time streaming data [6,7]. Data assets include real-time and near real-time data-streams from sensor networks, web services and city infrastructures.

Big data from emerging dispersed heterogeneous environmental sensing networks can include a broad range of data types and sources. They are volunteered and crowd-sourced citizen data from social media, mobile apps and citizen participation/citizen science platforms, and also traditional static historic/legacy data sources. Data can take many forms, from conventional data stored in Relational Database Management Systems to file and document/blob stores (or 'lakes') for unstructured data and/or real-time streaming data being generated by IoT sensor networks, satellites [8,9], and web Application Programming Interfaces (APIs) [10]. Such data can be generated from a wide range of sources – which are increasing in availability, much of it being open source in origin [11]. The open source philosophy has revolutionised data-driven approaches and techniques. It has been shown that big data approaches offer the potential to improve environmental modelling and urban analytical approaches, including assessment of performance of urban regions and smart city infrastructures [12], and also in applications where predictive or real time environmental assessment is required – for example in flood preparedness and coastal management [13], risk analysis [11,14], or in the environmental impacts of pollution and contamination [15].

2.3.1. Network data flows

Data transfer flows within and between networks have three broad aims, outputting purposeful data (or 'capta') (e.g., ambient temperature), ingesting external operational data (e.g., enabling management of operation of the network and its components based on external analytics) and internally exchanging operational data (i.e., enabling management of operation of the network and its components based on internal decision trees). These data flows can be on a range of scales in terms of content, timing and extent (e.g., geographical location). The latter two flows are primarily part of network management (i.e., system maintenance) and the former is focused on analytics and monitoring and is usually the main flow in scale and the usually underpins the fundamental purpose of the network. Data governance rules are vital to these kinds of analytics networks and can be broadly grouped into 6 categories (Fig. 2).

1. **Data Ownership.** Data ownership can be a complex issue especially where the network has multiple users and sponsors (e.g., a science-driven network installation supported through government operational permissions), and data fusion is used in the pre-processing of data streams [16]. Definition of ownership and Intellectual Property (IP), at the network design stage is also vital, as are planning exercises for downstream analytics and/or product combination.

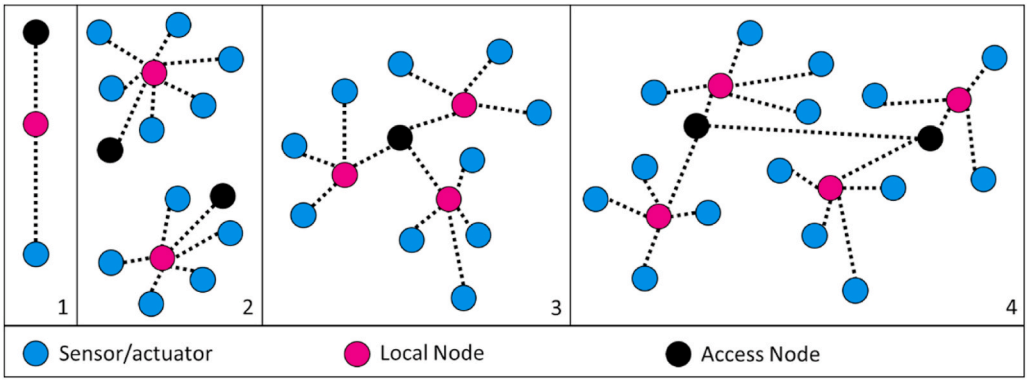

Fig. 1. High level schematic of dispersed topologies. Partially presented in a data flow view and partially an implementation view. Configuration from Left to right: 1). Direct sensor, configuration; 2) Dispersed node network, configuration (with two netwoks shown); 3) Connected node network and configuration; 4) Mesh network [24].

1. **Direct sensors**: This topology can be permanent or switched. A permanent direct sensor topology is a hardwired or fixed wireless connection established between two points. Switching is based on a connection that can be moved between different access nodes. This topology type is commonly used in applications where a single sensor is used together with a smartphone, tablet or dedicated microprocessor acting as a data aggregator.
2. **Dispersed node network**: Each network of this type consists of a single "central node," such as a hub or a switch that every sensor in the local network connects to and a dedicated access node. This topology is easy to design, implement, and extend. All data traffic flows through the network central node, thus an appropriately intelligent central node is required. Any failure of this central node will result in the consequent loss of the entire local network, so contingencies may also be required in the design. The dispersed node network topology is one of the most common sensor network topologies. A wireless personal area network, consisting of a gateway connected to several wireless sensors, is an example of this topology.
3. **Connected node network**: For this topology there is a single access point to the network 'tree', used either for direct data access or data telemetry. This access node is connected remotely to a series of separate local nodes with their associated sensors/actuators. This is a hierarchy of nodes in which the highest level is a single "root node," and this node is connected to one or many nodes in the lower level. Such a topology can contain many levels of nodes. The processing and power in nodes increase as the data moves from the branches of the tree toward the root node, allowing data to be processed close to where it is generated (an 'edge computing' approach). This topology is scalable, and its simple structure makes it easy to identify and isolate faults. Failure will not necessarily compromise the entire network. However, tree networks become increasingly difficult to manage as they become larger, and network edges are used as gateways.
4. **Mesh network:** For this topology there are multiple access nodes, and the network has the potential for bi-directional intra and extra network communication. In this approach the nodes disseminate their own data and also act as relays to propagate the data from other connected nodes. There are two forms of mesh topology: a partially connected mesh, in which some nodes are connected to more than one other node, and a fully connected mesh, in which every node is connected to every other node in the mesh. Mesh networks are resilient and self-healing, as data can be routed along a different path if a node fails. However, fully connected mesh networks are not suitable for large sensor networks as the number of connections required becomes unmanageable. Partially connected mesh networks provide the self-healing capability of a fully connected network without the connection overhead. Mesh topologies are most commonly found in wireless networking based on the IEEE 802.15.4 standard and its derivatives [4]. This is potentially the most advanced network type as it allows an integrated dispersed network control system to be implemented.

2. **Transfer Protocols.** Transfer protocols are concerned about data security, International Organization for Standardization (ISO) standards, ethics, interagency access and international streaming. For emerging dispersed heterogeneous environmental sensing networks, transfer protocols can be used to embed measures for data security such as checksum verification and content validation, whether or not ISO standards (e.g., number and function) are required depends on the type of sensing network being used and the target user groups.

   Ethical considerations around data sharing need to be assessed. These considerations could be associated with the compilation and storage of personal data (either as a primary goal or as a secondary effect, e.g., hospital admissions or people's movements) or ambient data. The ethics of data transfer to other agencies, especially those with statutory responsibilities or powers also need be considered given that some measurements once processed and validated have legal and compliance implications (e.g., mixing ratios of a regulated pollutant gas species). The transfer of data in this way also has data ownership implications. Similarly, for international data transfer, additional national governmental jurisdictions and considerations need to be accounted for, as some data may be considered a national asset, security issue or saleable item. Hence, it needs specific permissions for onward sharing or dissemination.
3. **User Interfaces:** The level of data presented, its format and level of processing are defined by the user group(s) it is aimed at. The differences between science and engineering requirements and government and public users and uses may be significant (e.g. tabulated sensor data output vs graphical representation of information). In many cases direct output data is relatively meaningless without processing (e.g., transformation from initial machine counts, hexadecimal compressed files or from raw voltages). Once data is in a human readable form and presented in understandable values it may not be suitable for general dissemination (uncorrected, uncalibrated or not yet validated) and may require further transformation depending on the complexity of the user requirement. For certain types of sensors or actuators, these transformations can be undertaken at the node (edge) itself and the data simplified before onward transmission.
4. **Compliance Implications.** In cases where there are jurisdictional compliance requirements or implications, data needs to be validated and calibrated appropriately. In parallel, submission protocols need to be established for creating regulatory data products.
5. **Data Security.** Operational backups (including redundant systems) are important for supporting long term data security. Depending on network requirements these measures can act upon varying levels of processed data. The data may be secured with controlled access or be anonymised depending on ethical considerations. Particular care needs to be taken in regard to multi-party cloud-based data access, with different products having different levels of security depending on product and user. It is important that appropriate care should be taken to consider the security at sensor (e.g., locally stored or held in memory for transmission or processing).
6. **Data Legacy.** Plans for data legacy need to be incorporated into network design. This includes IP control (related to the data ownership) of both the pre-processed data and the various levels of processed data. In many applications best practice would retain both the original data and the mechanism for its processing and transforming. In other applications synthesised products with contextual data are the appropriate products to be stored. This data needs to be

both a faithful record from the network and the analytics approaches used as well as contains the needed (as defined by user groups) useful data. Associated with data legacy is the importance of maintaining appropriate levels of metadata being generated so as to enable discovery and findability of data assets. Various international ISOs and related standards may be adopted, e.g., ISO19115 [17], Dublin Core [18], Data Catalog Vocabulary - DCAT [19], alongside national equivalent forms such as the UK GEMINI [20]. The notion of 'FAIR' data principles (Findability, Accessibility, Interoperability, and Reusability) has emerged as an axiom describing best practices for data producers and consumers [21].

2.3.2. Communication and data exchange

Suitable mechanisms are required to stream, capture and provide data analysis. A current commonly used example of such a platform is Microsoft Azure. A typical processing chain takes captured data, aggregates it and integrates it within a database. For Azure, this means routing sensor data to the cloud based 'IoT Hub', which can handle millions of messages from environmental sensors. Once received in the IoT Hub, Azure's 'Data Factory' permits the data to be transformed (with data extract, transform and load processes), to be routed to different recipient applications e.g., data stores (NoSQL data stores such as blob and document file stores and data lakes, traditional SQL table stores and specialised stores), analytical environments (e.g., offline analytics, online routine systems or the Microsoft Azure 'Machine Learning Studio') or to visualisation tools (such as the Microsoft 'PowerBI dashboard', able to generate web or mobile reports and event-driven push notifications).

3. Generalised network architecture

A schematic overview representation of the options within a dispersed heterogeneous environmental sensing network is presented in Fig. 3. This generalised network architecture model highlights some of the broad issues that need to be defined and addressed across a wide range of network topologies. This model can be configured in a number of ways and can be altered to suit the type of installation. It can also be altered during the life of the network (the network installation, expansion and network draw down phases for example would each have different requirements). Fig. 3 identifies the top-level pathway and options from sensors/actuators to options for client facing interfaces via local and/or cloud analysis and storage, application frameworks and service oriental architectures. It also shows the underlying considerations that need to be addressed (e.g., data security, governance and maintenance). Elements of this flexible generalised network model can in principle be implemented for a range of specific applications and three examples of this are presented here. Specifically, the "Reinvent the Toilet Challenge" (RTTC) and the Urban Observatory and Living Laboratory initiatives at Cranfield University.

4. Case studies

The case studies presented here are selected to illustrate key elements of the generalised network architecture model. The first case study is a physical analytics test kit developed as part of the RTTC. The original purpose of the RTTC was to capture data during field testing for operational and functional feedback. As the challenge developed, this purpose was expanded to include feedback-based network control and the capture of in-situ environmental data. The aim is for the RTTC to develop into a combined sensing, control and analytics network system suitable for a range of purposes. The second case study is the Cranfield Urban Observatory (https://www.livinglab.ac.uk). The Cranfield Urban Observatory is part of a network of UK Urban Observatories funded through the UK Collaboratorium for Research on Infrastructure and Cities (UKCRIC) initiative. This network of networks builds on elements of the RTTC and is designed to monitor and characterise a wide variety of environmental factors, infrastructure systems and behaviours (e.g., water distribution, power grids, biodiversity, environmental data, street and pedestrian behaviour). The third case study considers the role of cloud-based storage and analytical platforms and the value of the providers for projects such as the Cranfield Urban Observatory, taking the UKCRIC 'Data and Analytics Facility for National Infrastructure' (DAFNI) platform as an example.

4.1. RTTC analytics sensor test kit

The Gates Foundation RTTC project aims to develop a solution for sanitation issues in low-income countries. A physical analytics test kit was developed as part of this project for monitoring and advanced process control suitable for a broad range of environments. The RTTC analytics framework is based on a mixture of network topologies to allow flexibility of use in the field. The primary aim of this analytics test kit is to capture data on RTTC field testing progress and to make this data

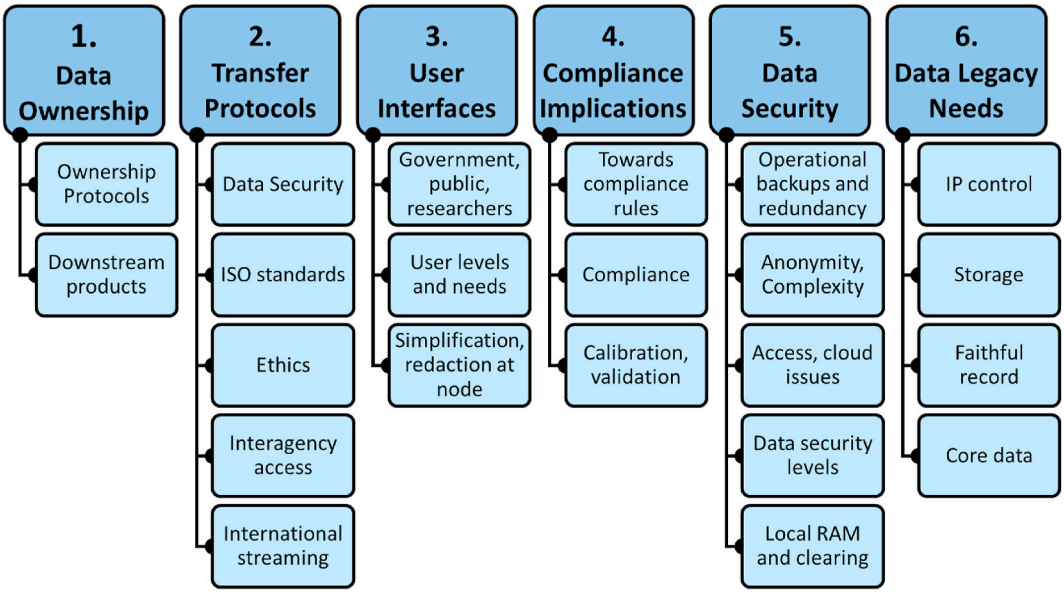

Fig. 2. Overview of data governance considerations for emerging dispersed heterogeneous environmental sensing networks (excluding analysis).

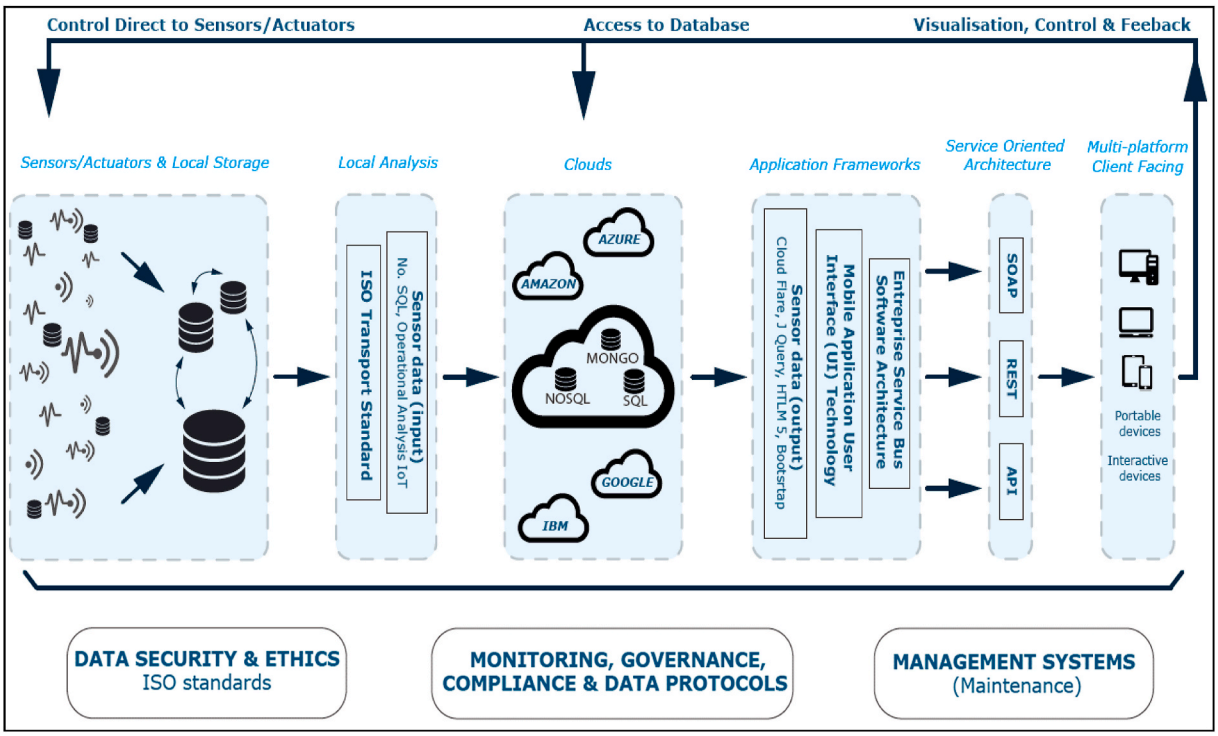

Fig. 3. Dispersed Heterogeneous Environmental Sensing Network architecture (adapted from Ref. [24]).

available for remote access, analysis and presentation. This scope has expanded beyond immediate operational and functional analytics to include wider environmental data associated with a range of water, sanitation and hygiene initiatives [22]. This broadening of scope is to allow more integrated installation and improved operational control for each specific installation environment.

The RTTC is based on development and adoption of sanitation technologies, suited for use in developing countries. These technologies are aimed to be independent of infrastructure for water and sewage as well as transforming collected waste into a useable product. The RTTC analytics sensor test kit was developed to generate and remotely capture operational and environmental data for RTTC field testing. The main parameters for the test kit are pH (used in the main RTTC reactor to trigger an internal water pump) and temperature (across the system e.g., trigger power use as well as safety features). The RTTC analytics sensor test kit has been designed to be physically robust and affordable with an associated cloud-based data support structure. The RTTC project infrastructure was used to provide a trial input environment for the analytics test kit. The kit was used for advanced RTTC process control and data collection during field-testing. The primary aim of the analytics test kit is to capture data on field testing progress, to enable remote feedback driven control of the RTTC (for example power use or water circulation), and to make data on the RTTC and its wider environment available for remote access, analysis and presentation. This scope has expanded beyond operational and functional analytics (pH and T) to include wider environmental data associated with a broad range of water sanitation and hygiene initiatives (e.g., footfall, precipitation, environmental temperature, atmospheric pressure, air quality, etc.). The ultimate aim is to generate a complete analytics system and an end-to-end approach suitable for adaptation to other infrastructures or environments in need of monitoring and observation. Integral to this work is that this system remains consistent with the Gates Foundation Open Access Policy [23]). The analytics sensor test kit is an open landscape development tool which is target-driven, with a focus on system management, and production of data analytics and functional outputs (e.g., the RTTC reactor control and power management). The analytics sensor test kit approach is designed to enable development of systems evaluate and manage a range of installations and systems in real-world use.

There are three primary components to the RTTC analytics sensor test kit management framework: First is "*operation*", being related to the sensors, instruments or machines used for implementation; Second is "*environment*", referring primarily to the environmental media of the operational component (e.g., soil, air, people); Third is "*communication*", relating to data flows and transport, signal fidelity, security and outputs (Fig. 4). Communication has an overarching role in network operation. This end-to-end system is designed to be open access and to provide output data products freely available to interested stakeholders.

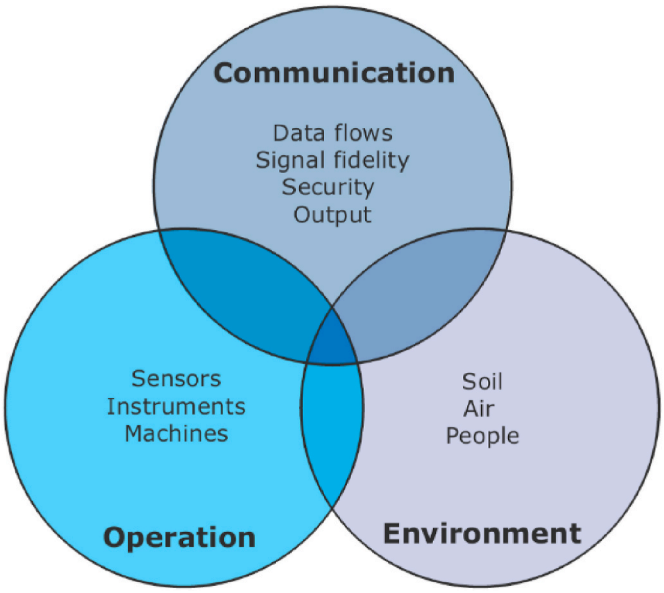

Fig. 4. Management framework for the RTTC analytics sensor test kit showing key points for each aspect of the framework.

In addition, the purpose of the analytics test kit is to support the development of broad architecture capable of measurement, data assimilation and system communication with real-time, whole system process control during testing and operation of the RTTC. Fig. 5 shows a particular case schematic used as part of the RTTC project, being namely the "low-cost monitor and control analytic IoT test kit" used in the development and testing phases of the RTTC project at the project test site in Birla (India).

The two main purposes of the sensors within the Birla RTTC network implementation are to: i). provide data for routine system control (via network actuators) and ii). provide information on the wider environment in which the network is located. Feedback to actuators can be automated either directly at the sensor itself (e.g., embedded operational firmware control) or from in-network control via the local nodes (if the sensor and local node are combined then these are in essence the same). The RTTC is geared towards the use of high numbers of low-cost sensors, located alongside a lower number of actuators and direct sensor inputs. Feedback to control actuators can be based on changes in network-sensed parameters, according to decision trees or lookup tables, each of which can be pre-defined (e.g., in the Birla installation, to increase water flow or change setting in the electro chemical reactor, see Fig. 5, left panel). Feedback can be provided in-network, based on the local node, or in dispersed mesh networks from remote in-network nodes. Feedback from sub elements of the network can be used to change overall network behaviour or to change behaviour for selected parts of the network. For example, information from a network component remote from the system under control (e.g., a change in weather in one part of the network changing the RTTC drainage control across the network or just in the path of the weather change). Remote automated feedback can be also be input to the access node and therefore onwards to the network, by user control direct to the sensors, by actuators (e.g., locally connected laptop), or via telemetered input from remote users or off-line analysis systems.

Data reporting between sensors is inherently very diverse, e.g., a simple analogue output within a defined voltage range, or complex tabulated data with associated operational "housekeeping" and calibration or quality meta-information. Rates of data acquisition are also variable within the RTTC. Data needs to be stored locally (even if briefly) in some form before being made available via the access node as a rolling backup to prevent catastrophic information loss. The data storage format needs to be robust as at this initial stage there is no backup to the data being collected (data security is discussed in 2.3.1). The data generated by the components of the analytics system is often initially in a machine-readable form which is unconverted into a human

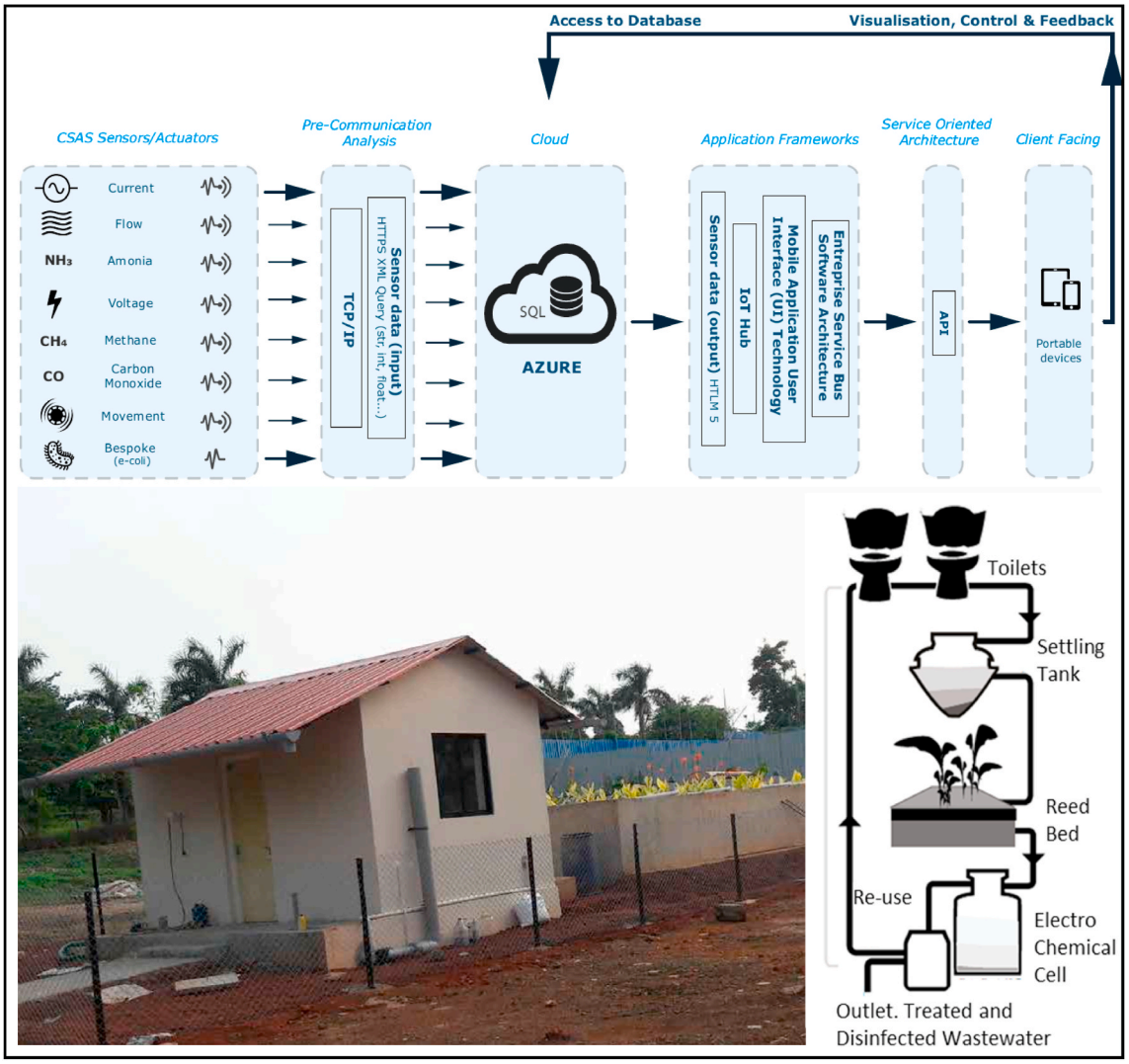

Fig. 5. (Top) RTTC project "Low-cost monitor and control analytic IoT test kit" test architecture. (Bottom left) Birla Institute of Technology and Science field tests site. (Bottom right) Birla site schematic showing from top to bottom the toilets, settling tank, reed bed, electrochemical processing cell and processed wastewater outlet (adapted from Ref. [24]).

readable form or format. This is usually a function of the sensor operation itself (e.g., machine counts representing a capacitance value within a set range which is proportional to a humidity value). This data is then transmitted to the local node where it can be managed. Management operations can involve longer term storage, transformation (e.g., to human readable data or conversion using pre-set conversion tables), compilation (with input from other sensors/actuators of both similar and dissimilar types) and compression (for efficient long-term storage, preparation for access or telemetering to preserve bandwidth). This managed data can then be made available for local access or telemetry.

The RTTC local data reservoirs (Fig. 5) have variable storage requirements and are flexible enough to enable linkages between a range of network sensor and actuator types as well as network distributions and feedback linkages (e.g., temperature within the RTTC body will be treated differently to temperature observations outside the RTTC installation block). Communications between devices and nodes use Message Queue Telemetry Transport. The RTTC implementation is discussed in detail in Williams et al. [24]. The purpose of this layer is to generate relatively lightweight combined messaging products for the local and or access nodes (where conventionally they would remain separate).

The RTTC architecture is designed for local and remote access via the access nodes (automated or non-routine). Data transmission can be bi-directional, but the highest volumes of traffic will be from the sensor data local archives. This is considered to be network "Input" sensor data (Fig. 1). Telemetered communications from the RTTC access node will be based on the cloud-based NoSQL architecture, which is powerful, widely used and robust, holding documents at an entity level. Direct user control can also be used to access the stored cloud data. At this level the data can also be further managed with a combination of transformation, compilation and compression tasks.

4.2. Cranfield Urban Observatory and Living Laboratory

4.2.1. Introduction to Urban Observatories

The types of dispersed heterogeneous sensing networks described, and employed in the RTTC, together with the increasing recognition of the need to understand complex urban systems in the face of contemporary challenges (e.g., climate change, sustainability, resilience), are helping to drive the development of Urban Observatories. The term 'Urban Observatory' itself is not new (e.g., Ref. [25], but previously has been commonly focused on single discipline studies, lacking the measurement, integration and analysis of data spanning the whole urban system and the Five Capitals, namely Natural, Human, Social, Built/-manufactured and Financial [26]. However, recent developments in low cost IoT based dispersed heterogeneous environmental, infrastructure and social sensor configurations, are driving a growing interest in the creation of Urban Observatory systems to capture, analyse and present urban system behaviour across the Five Capitals.

4.2.2. Cranfield University Urban Observatory

The Cranfield University Urban Observatory, which is currently under development, is designed to allow users to monitor, characterise and compare a wide variety of environmental and infrastructure systems and behaviours. These currently include (but are not limited to): water distribution, power grids, biodiversity, pollution, environmental data, street networks, population density, biodiversity and pedestrian behaviour. The experience gained through the RTTC project is now being used to underpin the development of the Cranfield University Urban Observatory, which forms part of a wider network of six Urban Observatories in the UK funded through the UKCRIC [27]. The Cranfield University Urban Observatory is unique, being located on a self-contained site in a semi-rural location, representing a microcosm of a complex city system. The site not only includes traditional campus infrastructure (e.g., halls of residence, teaching and research facilities) but also complex infrastructure including an airport, combined heat and power plant, solar farm and wastewater experimental and treatment works for example.

A recurrent theme in this Observatory is the adoption of the 'digital twin' as a design approach designed to create a 'digital analogue' of the systems in contention. The digital twin draws on the modal themes described in the RTCC design approach, using environmental sensors to capture facets of a complex system in real, or near-real time. This permits extenuated management decisions to be adopted and trialled as scenarios, as well as recreation of circumstances leading up to particular events. A fundamental aspect of the Observatory is that the data arising from these assets will be brought together, using a common data platform, within the envelope of the Cranfield University Urban Observatory, and accompanied by a diverse range of analytical and dissemination tools and functionality. The Observatory project team identified these fundamental aspects as being critical to enabling research and generating knowledge. In addition, the Observatory has been designed to enable data sharing, for example with the wider UKCRIC Urban Observatories, and data users, via common data structures and APIs.

In parallel with the increased interest in Urban Observatories has been the development of the concept of 'Living Laboratories'. Whilst Living Laboratories are not Urban Observatories *per se*, they are increasingly the mechanism by which research questions are framed, facilitating interaction between stakeholders, students, academics and Professional Service Units (e.g., campus facilities), allowing any of these groups to generate research questions [28,29]. The Cranfield University Urban Observatory forms an integral part of the wider Cranfield Living Laboratory initiative. This is developing the use of the Cranfield University campus as a testbed for transformative technologies and approaches for the delivery of enhanced social, economic and environmental outcomes in response to major challenges such as future urban design, hybrid transport systems, provision of sanitation, well-being and a range of issues associated with integrated infrastructure development. Here the data provided by the Urban Observatory is facilitating academic research, teaching, and operational decision-making in relation to such issues.

Whilst still under development, the Urban Observatory and Living Laboratory initiatives at Cranfield University have had to grapple with many of the technical issues highlighted in this paper. In particular, the Urban Observatory seeks to integrate data from a wide array of differing sensor and data types, often communicating using differing protocols, data formats/structures and at widely varying volumes, and veracity. Experiences to-date suggest that identifying a 'one-size-fits-all' solution may prove challenging. Whilst the Urban Observatory has highlighted technical challenges, the Living Laboratory is highlighting wider emerging issues, particularly with regard to ethics and governance, which will assume increased importance as the deployment of dispersed heterogeneous environmental sensing networks, Urban Observatories, and Living Laboratories accelerates.

4.3. Data and Analytics Facility for National Infrastructure

A key to the success of IoT installations, such as the Urban Observatories, and Living Laboratories described is the ability to collate together in real or near real time the multitudinous data feeds arising at scale from the many environmental sensors in place, and the ability to present this data efficiently to analytical and visualisation tools. There are a number of pertinent challenges associated with this. Firstly, the volume, variety and velocity of the data arising require specialised hardware, network and data transport configurations. Secondly the bodies of data so collected need to be made accessible for further processing, manipulation and output. Finally, in an information system seeking to explore and model potential scenario outcomes, for example the consequence of meteorological phenomena such as temperature variations on pedestrian flows, data from a range of differing sources

may need to be bought together and co-consulted. The DAFNI [30] operated by the UKRI Science and Technology Funding Council is one such 'High Performance Computing' platform able to hold hundreds of substantial 'legacy' data volumes, concerning themes such as infrastructure networks, socio-demographic profiles and cadastral information. Combining this capability with more real-time sources of infrastructure and environmental data opens a route to data-informed decision support tools.

Once the data is received it needs to be held in a manner whereby the Digital Twinning modelling approaches adopted can permit a decoupling from the observed system, the tool being used to inform maintenance and/or design for the future operation of the real system; the digital twin thus being run offline to permit improvements to the real system [31]. The DAFNI facility is being designed therefore to ingest substantive real-time sensor data to be placed alongside extant static data volumes, with all these being then made directly available to the suites of modelling and visualisation tools. The complexities as to which datasets should be interoperated together are forming the focus of an ongoing development to establish 'infrastructure research ontologies' (IROs), linked data themes that in concert can be used to support infrastructure modelling challenges, for example optimal and demand-driven siting of plant and infrastructure facilities. The IRO approach will facilitate operators being able to determine which data themes are appropriate for inclusion in a modelling approach. However, another challenge is the sheer volume and diversity of the data held and therefore, connected with this, will be the need to enable efficient data discovery through associated metadata records, drawing on standards such as DCAT [19]. Through the combination of these approaches, DAFNI aims to improve the efficiency, reliability and sustainability of infrastructure through better sharing and use of data, the exploitation of simulation and optimization techniques, and engagement with stakeholders through visualisation.

5. Conclusions and discussion

The DEE, comprising a sensor architecture with a broad measurement and actuation capability, coupled with a cloud-based computing approach is capable of assimilating environmental and human user inputs in real time, producing network relevant outputs for monitoring and managing environmental states. In this paper, we have explored different case studies to show advances in the field. With the RTTC analytics test kit, progress to date has identified a number of areas of broad relevance to network design (taken to mean the selection of network components, their physical layout and their mode of operation) as well as corresponding potential solutions. The RTTC outcomes provide a demonstration of this type of 'whole system of systems' process control approach for networks up to and including dispersed mesh approaches. Many of the lessons learnt during the development of the RTCC analytics test kit are more broadly applicable and will influence current activities such as the ongoing evolution of the Cranfield University Living Laboratory.

The Cranfield University Urban Observatory extends the scope of the RTTC in the laboratory context of a functioning and characterised campus/semi urban environment. The RTTC framework will be deployed as part of the Cranfield Urban Observatory and so will be integrated with a wide variety of other Urban Observatory applications with the aim of exploring avenues for potential inclusion of viable and valuable extension of the RTTC in the field. The quality of low-cost sensors is improving rapidly, and many are already suitable for applications such indoor and outdoor air quality [32]. Their greater flexibility and lower cost mean that they can be deployed in a ubiquitous manner, and used to address many more questions than conventional, more expensive instruments could, such as those which were developed specifically to meet regulatory needs. The corresponding rise in complexity in new sources of data, derived in part from associated reduction in the costs of both the generation and storage of data over the last decade, has in turn led to the creation, and growing ease of access to substantive datasets. Corresponding predictive analytics, machine learning and stream analytics are required to mine and reveal patterns and clusters in voluminous, fast-changing, diverse, structured and unstructured data sources to develop data intelligence. These big data approaches offer a wide range of opportunities by embracing and extending traditional informatics approaches, allowing a level of data processing that would traditionally have been unachievable [33] and yet which are now becoming increasingly necessary. Facilities such as the DAFNI can then be utilised to enable the data arising from multitudinous networks of infrastructure and environmental sensors. They can be deployed alongside substantive 'data lakes' of extant data, and to thence be made available to the advanced analytical and visualisation tools. The tools are used to observe complex systems, investigate the impact of particular mitigation measures, and have the potential to provide a basis for understanding previous patterns of outcomes through 'hindcasting' approaches, offer near-real-time information and feedback systems through 'nowcasting' approaches, as well as supporting exploratory future scenario modelling through 'futurecasting' approaches.

Credit author statement

The authors have confirmed their equal contribution in this manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to thank the Bill & Melinda Gates Foundation and the wider RTTC team at Cranfield University and at partner Institutes. The authors would further wish to thank the Cranfield Water Science Institute laboratories, the Cranfield Centre for Competitive Creative Design, the Cranfield Energy Centre and the Cranfield Centre for Atmospheric Informatics and Emissions Technology. The Cranfield University Urban Observatory is funded by the UKCRIC: Urban Observatories (Strand B) – (EPSRC EP/P016782/1) and the UKCRIC – CORONA: City Observatory Research platfOrm for iNnovation and Analytics project (EPSRC EP/R013411/1). These together with the UKCRIC National Infrastructure Database, Modelling, Simulation and Visualisation Facilities (EPSRC EP/R012202/1) are supported by the UKCRIC Coordination Node (EPSRC EP/R017727/1), which funds UKCRIC's ongoing coordination.

References

[1] Raffaelli D, Bullock J, Cinderby S, Durant I, Emmet B, Harris J, Hicks K, Oliver T, Paterson D, White P. Big data and ecosystem research programmes. Adv Ecol Res 2014:51–77.
[2] Pollard JA, Spencer T, Jude S. Big Data Approaches for coastal flood risk assessment and emergency planning. WIREs Clim Change 2018;9:e543.
[3] ISO. Information technology — open systems interconnection — basic reference model: the basic model. ISO/IEC 7498-1:1994. https://www.iso.org/standard/20269.html. [Accessed February 2022].
[4] IEEE. IEEE standard for low-rate wireless networks. IEEE 802.15.4-2020. https://standards.ieee.org/standard/802_15_4-2020.html. [Accessed February 2022].
[5] Fujitsu Ltd.. White paper linked data: connecting and exploiting big data. March 2012 ref: 3378, by I. Mitchell and M. Wilson. Fujitsu Services Limited; 2012. p. 1–21. 2012.
[6] Jagadish HV. Big data and science: myths and reality. Big Data Res 2015;2:49–52.
[7] Marr B. Big data: using SMART big data, analytics and metrics to make better decisions and improve performance. Cichester: Wiley; 2015258pp.
[8] Maier D, Megler VM, Baptista AM, Jaramillo A, Seaton C, Turner PJ. Navigating oceans of data. Scientific and Statistical Database Management Conference; 2012. p. 1–19.
[9] Moszynski M, Kulawiak M, Chybicki A, Bruniecki K, Bieliński T, Łubniewski Z, Stepnowski A. Innovative web-based geographic information system for municipal

areas and coastal zone security and threat monitoring using EO satellite data. Mar Geodes 2015;38:203–24.
[10] Singhal A, Pant R, Sinha P. AlertMix: a big data platform for multi-source streaming data. https://arxiv.org/ftp/arxiv/papers/1806/1806.10037.pdf. [Accessed February 2022].
[11] Rumson AG, Hallett SH. Opening up the coast. Ocean Coast Manag 2018;160:133–45. https://doi.org/10.1016/j.ocecoaman.2018.04.015.
[12] Caird SP, Hallett SH. Towards evaluation design for smart city development. J Urban Des 2018;24(2):188–209. https://doi.org/10.1080/13574809.2018.1469402.
[13] Rumson AG, Hallett SH, Brewer T. Coastal risk adaptation: the potential role of accessible geospatial Big Data. J Mar Pol 2017;83:100–10. https://doi.org/10.1016/j.marpol.2017.05.032.
[14] Choi T-M, Lambert JH. Advances in risk analysis with big data. Risk Anal 2017;37:1435–42.
[15] Romero JM, Hallett SH, Jude S. Leveraging big data tools and technologies: addressing the challenges of the water quality sector. Sustainability 2017;9:2160. https://doi.org/10.3390/su9122160.
[16] Castanedo F. A review of data fusion techniques. Sci World J 2013;2013(704504):19. https://doi.org/10.1155/2013/704504.
[17] ISO. Geographic information — metadata — Part 1: fundamentals. ISO 19115-1:2014. https://www.iso.org/standard/26020.html. [Accessed February 2022].
[18] ISO. Information and documentation — the Dublin Core metadata element set — Part 1: Core elements. ISO 15836-1:2017. https://www.iso.org/standard/71339.html. [Accessed February 2022].
[19] W3C. Data catalog vocabulary (DCAT) - version 2. https://www.w3.org/TR/vocab-dcat-2. [Accessed February 2022].
[20] AGI. UK GEMINI standard and INSPIRE implementing rules. https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini/40-gemini/1037-uk-gemini-standard-and-inspire-implementing-rules. [Accessed February 2022].
[21] Wilkinson M, Dumontier M, Aalbersberg I, et al. The FAIR Guiding Principles for scientific data management and stewardship. Nat Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
[22] UNICEF. Water, sanitation and hygiene. https://www.unicef.org/wash/3942_3952.html. [Accessed February 2022].
[23] Gates Foundation. Bill & Melinda Gates foundation open access policy. https://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy. [Accessed February 2022].
[24] Williams, et al. Gates foundation reinvent the toilet challenge (RTTC) analytics sensor test-kit for advanced process control. 2018. Technical report OPP1164132 (in press).
[25] Jones V. Urban observatory programme. Urban Aff Q 1972;8:35–9.
[26] Maack M, Davidsdottir B. Five capital impact assessment: appraisal framework based on theory of sustainable well-being. Renew Sustain Energy Rev 2015;50:1338–51.
[27] UKCRIC. UK collaboratorium for research on infrastructure and Cities. https://www.ukcric.com. [Accessed February 2022].
[28] Westerlund M, Leminen S, Habib C. Key constructs and a definition of living labs as innovation platforms. Technol Innovat Manag Rev 2018;8:51–62.
[29] Westerlund M, Leminen S, Rajahonka M. A topic modelling analysis of living labs research. Technol Innovat Manag Rev 2018;8:41–50.
[30] Hall J. UK reveals new platform for infrastructure data analysis and simulation modelling. 3. In: Proceedings of the institution of civil engineers - civil engineering 2019, vol. 172; 2019. p. 102. https://doi.org/10.1680/jcien.2019.172.3.102. 102.
[31] Batty M. Digital twins (editorial). Environ Plann B: Urban Anal City Sci 2018;45(5):817–20. https://doi.org/10.1177/2399808318796416.
[32] Mead MI, Popoola OAM, Stewart GB, Landshoff P, Calleja M, Hayes M, Baldovi JJ, Hodgson TF, Mcleod MW, Dicks J, Lewis A, Cohen J, Baron R, Saffell JR, Jones RL. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmos Environ 2013;70:186–203.
[33] Vitolo C, Elkhatib Y, Reusser D, Macleod CJA, Buytaert W. Web technologies for environmental big data. Environ Model Software 2015;63:185–98.