OPEN DATA OBSERVATORIES

Naeima Hamed
School of Computer Science and Informatics
Cardiff University, UK
hamednh@cardiff.ac.uk

Omer Rana
School of Computer Science and Informatics
Cardiff University, UK
ranaof@cardiff.ac.uk

Pablo Orozco-terWengel
School of Biosciences
Cardiff University, UK
orozco-terwengelpa@cardiff.ac.uk

Benoît Goossens
School of Biosciences
Cardiff University, UK
goossensbr@cardiff.ac.uk

Charith Perera
School of Computer Science and Informatics
Cardiff University, UK
pererac@cardiff.ac.uk

September 5, 2022

ABSTRACT

Open Data Observatories refer to online platforms that provide real-time and historical data for a particular application context, e.g. urban/rural environments or a specific application domain. They are generally developed to facilitate collaboration within one or more communities through reusable data sets, analysis tools and interactive visualisations. Open Data Observatories collect and integrate various data from multiple disparate data sources – some providing mechanisms to support real-time data capture and ingest mechanisms. Data types can include sensor data (weather, traffic, pollution levels) and social media data. Data sources can include Open Data providers, interconnected devices, and services offered through the Internet of Things (IoT). The continually increasing volume and variety of such data require timely integration, management and analysis, yet presented in a way that end-users can easily understand. Data released for open access preserve their value and enable a more in-depth understanding of real-world choices. This survey investigates twelve active data observatories and the data that they provide. We provide a more in-depth analysis of six observatories established by the UK Collaboratorium for Research in Infrastructure and Cities (UKCRIC). An additional six observatories are then analysed based on their associations and shared concepts with the UKCRIC observatories, using different data management approaches. We investigated the aims, design and types of data used across multiple domains: transport, energy, environment and social sensing. We conclude with research challenges that influence the implementation of Open Data Observatories, outlining some pros and cons for each observatory and recommending areas for improvement. Our primary goal is to suggest best practices learnt from each observatory to aid the development of non-urban observatories.

Keywords Urban observatories, non-urban observatories, dashboards, data portal, smart city data, open data principles, linked open data, 5-star models

1 Introduction

Structured, semi-structured and unstructured data can be produced from different sources, including government authorities, academic institutions and citizens. Each source can use various methods to collect information, ranging from Internet of Things (IoT) devices to questionnaires and surveys. Many governments worldwide have published some of these data as Open Data – conversely, many commercial organisations also collect vast amounts of data, but only a small portion of these data is open [1]. Opening data can be achieved by using data observatories [2] that can include fine-grained raw data, and a repository of analysis techniques (e.g., statistical modelling, machine learning and artificial intelligence) to analyse and visualise such data [3, 4, 5, 2, 6]. Many existing observatories extract real-time data from IoT devices and transmit them to remote locations [7]. IoT devices can include sensors that collect observations from the source and interact with the associated controllers (consumer perspective) or a gateway (industrial perspective). Controllers aggregate streams of real-time data and transmit them to back-end systems such as IoT cloud platforms [8, 9]. Nevertheless, IoT cloud platforms serve as information repositories that enable data-centric actions such as modelling, analysis and visualisation. Legacy data systems, including data lakes and relational databases [10] are generally slow and siloed to cope with the ever-growing size and diversity of IoT data. However, Open Data Observatories can integrate, process and share these big and heterogeneous data in a timely manner, in addition to making them discoverable and accessible in a user-friendly format [2]. In the absence of competent data observatories, crucial information may lose value, become isolated and eventually become stale.

Our survey was inspired by Ma et al. [11] on finding timely solutions for managing IoT data across multiple smart city applications. Ma et al. predominantly aimed to bridge the gap between data collection and utilisation, surveying fourteen smart city data sets along with methods for data modelling and decision making.

Our survey reviews twelve data Observatories that collected, integrated and delivered real-time and historical data. We study key data management approaches from data generation, processing and presentation. Further, we highlight five challenges that may constraint their use and viability, such as integrating heterogeneous data while maintaining sufficient data quality, provenance and privacy. Our primary intention is to review existing literature to help researchers, developers, engineers and, stakeholders build urban and non-urban data observatories. More specifically, to suggest practical approaches learnt from each observatory to support the inferences required on how to develop data observatories effectively.

This survey is structured as follows: Section 2 investigates the use of the term Open Data. Section 3 introduces the twelve selected Open Data Observatories, individually describing their objectives, data management approaches and the (smart) services they support. Subsequently, suggests features that can be replicated in non-urban areas Section 4 recapitulates the types of data they support and provides insights into the modes of use for the reviewed observatories in domains of transport, environment, energy and social sensing. Section 5 describes and compares the data sources, formats, storage and processing approach for the reviewed observatories. This section also includes examples for applied predictive analytics and visualisation, explaining the employed techniques. Section 6 describes five key research challenges, namely data integration, context, quality, provenance and privacy based on our survey. We subsequently provide a critique of the reviewed observatories, suggesting future recommendations and scores based on the 5-star models. Finally, Section 7 summaries and concludes the survey.

2 Open Data

For the past decade, many individuals and businesses have used Open Data for analysis and software applications. In general, Open Data are non-personal, limitless, and free digital information [12]. Everyone can use Open Data as long as that they credit the sources [13, 14]. Predictably, Open Data are released in structured formats, accompanied by metadata, and presented in machine-readable formats [15]. For example, Spreadsheets (xlxs) [16], Comma Separated Value files (CSV), eXtensible Markup Language (XML), Javascript Object Notation (JSON), Shapefiles, Sequence (SP), Record Columnar (RC), Optimised Row Columnar (ORC), and Parquet files [17]. Machine-readable formats enable the computer's software to re-use, integrate and model the data for analysis. There are also a few inflexible Open Data formats, namely, Portable Document Format (PDF) and HyperText Markup Language (HTML), that computers cannot modulate directly. Open Data must satisfy the following criteria, as briefed in opendatahandbook.org and thoroughly discussed by Pereir et al. in [18].

- Available and accessible, the data must be complete, unaltered, and preferably downloadable over the internet in machine-readable formats.

- Re-use and re-distribute, the data must be permitted for full exploitation and re-publication, including merging with other datasets.

- Universal participation, the data must be non-discriminatory and non-restricted, equally offered to everyone.

Table 1: Description and comparison of the 5-star models for Open Data Forms [19] to support Open Data stakeholders in the technical section, Open Data Engagement [20] to recommend the engagement of end-users, and Open Data Portals [21] to guide all involved parties in building data portals.

Stars	Open Data Forms	Open Data Engagement	Open Data Portals
⋆	Portable Document Format (PDF) files.	Portal with external links to open datasets.	Portal with licensed Open Data.
⋆⋆	The above, plus spreadsheets (e.g. Microsoft Excel).	Add context and accurate meta-data to the above.	The above, plus structured and open meta-data.
⋆⋆⋆	All the above, plus comma-separated values (CSV).	All the above, plus seek users' feedback and reflect.	All the above, plus additional tools and codes for data re-use.
⋆⋆⋆⋆	All the above, plus semantic standards such as Resource Description Format (RDF).	All the above, plus build a network of skills. Encourage the public to re-use and analyse the data.	All the above, plus making portals the main data source with multiple formats that cater for a wider community of users.
⋆⋆⋆⋆⋆	All the above, plus linking data to external datasets (i.e., Linked Open Data (LOD))	All the above, plus work with other providers and involve citizens.	All the above, plus interoperable, offering open provenance, governance, quality metrics, and trust.

2.1 Open Data Usage and Benchmarks

Numerous Open Data applications exist in our everyday lives, but they provide both possibilities and difficulties. Online systems that interface with this data provide vital functions. During the worldwide spread of the infectious coronavirus (COVID-19) in December 2019, there was an immediate surge in demand for face masks and sanitising items, resulting in a retail supply deficit. Therefore, internet portals with real-time Open Data [22] assisted individuals in locating pharmacies that provide face masks. In contrast, the issues that may come from using Open Data generally concern the degree of real-timeliness, the quality standard, and the compliance with privacy regulations. For instance, the openness of real-time Open Data enables the provider to learn a great deal about the habits and lifestyles of residents. Consequently, it creates security and privacy concerns if the ethical use of data is not implemented with care [23, 22]. Multiple scientists and academics developed criteria to aid Open Data providers in putting their data online. Few recognised contributions were the 5-star Open Data models and associated hosting portals [21]. Sir Tim Berners-Lee, who built the web and launched the wheel of Linked Data, presented a 5-star model for Open Data of all sizes to assist Open Data stakeholders in the technical portion [19]. As such, Davies et al.[20] presented their 5-star approach to suggest end-user involvement. Colpaert et al. [21] developed a 5-star model for Open Data portals that may advise all stakeholders engaged in portal construction. Predominantly, the concept sought to improve data quality and encourage their reuse. Table 1 compares and contrasts the three 5-stars models and their respective descriptions.

2.2 Open Data Sources

Governmental agencies and academic institutions are crucial Open Data suppliers. They are only responsible for managing the technical and legal aspects of this data. The information made public may have originated from IoT devices deployed by several parties. The subject matter of data can range from science to the environment [26, 12]. A number of developed nations are required to publish their collected data. In the United Kingdom, for instance, the system had been in effect since 2009, when the government released a Command Paper committing to the public release of official datasets [27]. In 2010, the *data.gov.uk* website was launched to enable local authorities and public bodies to publish their data. These data representations, also known as data catalogues, contain datasets in numerous forms, such as CSV and JSON. Data Catalog Vocabulary (DCAT), a W3C recommendation, defined a dataset as a collection that has been collected and published by an organisation that permits access in many formats. Since its inception, the UK government's Open Data has grown tremendously, reaching over 40,000 datasets in November 2017 [28]. The primary purposes of such data are to promote transparency, re-use, improve public services, engage citizens, and create broader opportunities for innovations and best practices [29]. However, in 2011, Huijboom et al. [30] criticised the openness degree of the open government datasets in the UK. They randomly sampled 400 datasets from *data.gov.uk* and evaluated them using the eight original Sebastopol principles of Open Data [24]. Huijboom et al. verified that nearly two-thirds of published government resources are aggregated information instead of granular data, 38 percent are stale data, and 30 percent are inaccessible. Subsequently, the Sebastopol list was extended by Sunlight Foundation [25] to ten principles to enhance the openness and accessibility of government data. Whilst the matters of what data to open and how to open them remain ambiguous, some developed countries fear the consequences of opening all their data [23]. For example, Geospatial and Light Detection and Ranging (Lidar) data are one of the popular types as they contribute sustainably to official decisions on social and environmental matters [23, 31, 32, 33]. Besides, they can act

Table 2: Description and comparison of Open Data principles as proposed by Sebastopol [24], named after a meeting held in Sebastopol, California, in 2007 and gathered thirty open government advocates. The meeting agreed on eight principles for opening government data. In 2010, the Sunlight Foundation citeFoundation2010, a non-profit organisation that promotes open government, increased these principles to ten [25].

Principle	Description [25]	[24]	[25]
1. Complete	Datasets must be a complete and accurate representation of the original observations. Raw data and their meta-data must be unlocked, including all computations details. Decision-makers can exclude sensitive records that the Federal Law permitted their withholding (e.g. personally identifiable information).	✓	✓
2. Primary	Open governments datasets got collected at the source. Furthermore, they must include data collection methods and their supporting evidence (metadata).	✓	✓
3. Timely	Datasets must be published promptly after collection, especially the time-sensitive data, which may lose their value when disclosure delays. Real-time data are preferred in making accurate and informed decisions.	✓	✓
4. Accessible	Datasets must be easily accessible. Users should easily find the desired data, whether physically (by visiting official offices) or electronically (by downloading them from official online data portals).	✓	✓
5. Machine-processable	Datasets must be in a machine-friendly format. That is, the computer machine can process and manipulate them. For example, spreadsheets, CSV, XML and JSON.	✓	✓
6. Non-discriminatory	Everyone can access and use the governments' published datasets. Data acquisition is admissible without the need for registration, membership or even declaring the purpose of use.	✓	✓
7. Non-proprietary	Datasets must be in a freeware format. For example, a file in a Microsoft Office format costs money to use. Making the same file compatible with free software such as Apache OpenOffice grants access to a broader community of users.	✓	✓
8. License Free	Datasets must have a distinct label of public information that is freely available without restrictions or terms of conditions.	✓	✓
9. Permanence	Datasets must stay available online, stored in archives. In case of modification, all versions must be findable to enable users to track changes.		✓
10. Usage costs	Accessing and obtaining open government datasets must inquire no fees. Free and faithful data may encourage business growth and successively positively impact the overall economy.		✓

as a reference point for some of the modern and divisive technologies, including driverless vehicles and drones [23]. That said, Lidar data only came out publicly in a few countries, such as Finland in 2012, followed by Denmark and the Netherlands in 2013 and 2014, respectively [23].

3 Data Observatories

The term *observatories* derived from *observe* and referred to locations for monitoring territory. Originally employed by astronomers to study celestial objects with the aid of cameras and space telescopes. Similarly, data observatories are the web platforms that unify diverse data. They could appear under several names and titles. For example, dashboards [34, 35, 36], data portals [21], data platforms [37] and tracker project [38, 5, 39]. Urban data observatories provide accessible real-time and historical data. They, in turn, enable stakeholders and end-users to monitor the behaviour of cities and make informed decisions that may improve the performance of public services. Diversely, non-urban observatories [40, 41] focus on monitoring wildlife and generating data that support decisions to protect biodiversity from incidents, including forest fire and poaching. Building *urban data observatories* can be achieved timely in some cities due to the evolving network infrastructure, and the emerging modern network protocols such as ZigBee, Z-WAVE, INSTEON, WAVENIS, LoWPAN, NB-IoT [42]. Conversely, the *non-urban observatories* may require more communication work due to the lack of network infrastructure in remote areas (e.g., forest) [11, 43, 34, 44]. Throughout this section, we explore the twelve nominated data observatories - illustrating their systems design in figure 3 and suggest inspired ideas - summarised in tables 3 and 6 that could be potentially practical to replicate in the non-urban observatories.

3.1 Urban Observatory Project

One of the largest real-time environmental datasets providers worldwide is the Urban Observatory [45]. Sponsored by an integrated research capability resource named UK Collaboratorium for Research on Infrastructure and Cities

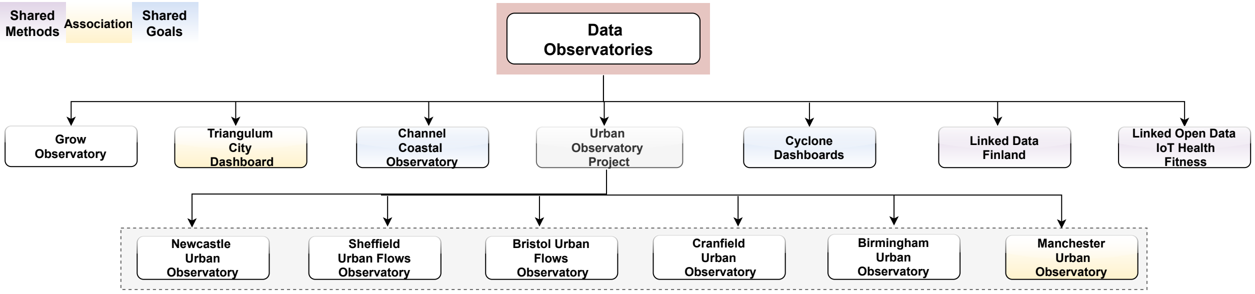

Shared Methods
Association
Shared Goals
Data Observatories
Grow Observatory
Triangulum City Dashboard
Channel Coastal Observatory
Urban Observatory Project
Cyclone Dashboards
Linked Data Finland
Linked Open Data IoT Health Fitness
Newcastle Urban Observatory
Sheffield Urban Flows Observatory
Bristol Urban Flows Observatory
Cranfield Urban Observatory
Birmingham Urban Observatory
Manchester Urban Observatory

Figure 1: Chart displays the reviewed Open Data Observatories

(UKCRIC) and led by Newcastle University. The project involves five more British universities (i.e., Sheffield, Bristol, Cranfield, Birmingham, and Manchester). They worked together to build observatories in their metropolitans. Each of which deployed a range of different sensors across its city to monitor the surroundings and record observations. The overall framework is distinct in applying scientific methods to support decision-making through multi-scale urban that observe, analyse, and model real-time and historical data. For example, air quality monitoring sensors deployed across Newcastle and Gateshead measure key air quality parameters such as Nitrogen Dioxide, Ozone, Carbon Monoxide and Particulates. These sensors generate accurate readings that both authorities and citizens can act upon them to reduce, for instance, exposure to air pollution [39]. There are over 50 data types, and many real-time datasets, freely available at the *www.urbanobservatory.ac.uk* website. These datasets compromise earth observations, traffic flow, air pollution readings, water quality parameters, and many more [45].

3.1.1 Newcastle Urban Observatory

Newcastle University leads the Urban Observatory project across the UK. It holds the world's most extensive collection of open sensing data [46]. Predominantly, the scheme mainly focuses on monitoring several urban indicators through IoT devices. It provides real-time and historical datasets such as traffic, vehicle statistics, weather, air quality, water quality, seismic signs, sewage monitoring, soil trace, noise detection, buildings' electric lights control, pedestrian count and many more [46]. The project has a large-scale of various smart devices capturing more than one hundred different metrics per second, besides static images, video, radar, and laser-scan matrices acquired separately. Currently, the system records over 7000 observations every minute from nearly 3600 active sensor streams, and 540 CCTV cameras [47]. Extracted data are published freely on the Newcastle Urban Observatory website [46]. Everyone can access and download the data - be it, researchers collecting datasets for experiments, policy-makers seeking evidence, citizens exploring the city and checking the weather, businesses accessing relevant, insightful information about the performance of demonstrated projects.

Technically, Newcastle Urban Observatory ingests streams of real-time observations in a cloud platform. The heterogeneous data pass through distributed file systems - client/server-based application that process data and instantly share them, simultaneously, on the local client's machine. For storage, MySQL and NoSQL databases served the purpose of storing structured and unstructured data, respectively [48, 17]. To cope with the collected data volume, velocity, and variety [49], employing Apache Kafka as a distributed messaging system was the chosen mechanism [50]. Kafka integrates the heterogeneous data for immediate sharing between different applications[48]. Nevertheless, a Representational State Transfer Application Programming Interface (RESTful API) enables researchers and developers to browse and access time-series data, locations, and even sensors. Everyone can leverage this API to integrate the observatory data into applications and use the downloadable CSV and JSON formats for analysis, modelling, and visualisations [46, 48]. Among the more of Newcastle Urban Observatory projects [51, 52, 53, 54, 1, 55], we took a closer look at the so-called Predicting Rainfall Events by Physical Analytics of Realtime Data (Flood-PREPARED) [56]. This project implements a pioneer resource for investigating real-time water surface flood risks and their effects on cities. It aimed to equip cities with novel physical, analytical methods to envisage surface water flooding and provide decision-makers with evident real-time predictions. The delivery of the project passed through five correlated stages as shown in figure 2. Another recent work by James et al. [57] who presented datasets quantifying the impact of COVID-19 measures in the UK. Existing IoT data and the fully-fledged analytics infrastructure enabled the authors to create an interactive COVID-19 dashboard. It visualises several indicators that update in real-time, comparing data changes with baselines. The dashboard also contains frequent automated comparative descriptive statistics (e.g. daily, weekly updates) to assist decision making [47]. For instance, observations gathered from air quality stations, car parks, and traffic sensors- when analysed- showed a steep drop in pedestrian footfall and traffic volume across Tyne and Wear city during the UK COVID-19 national lockdown in March 2020. Moreover, Newcastle Urban Observatory [46]

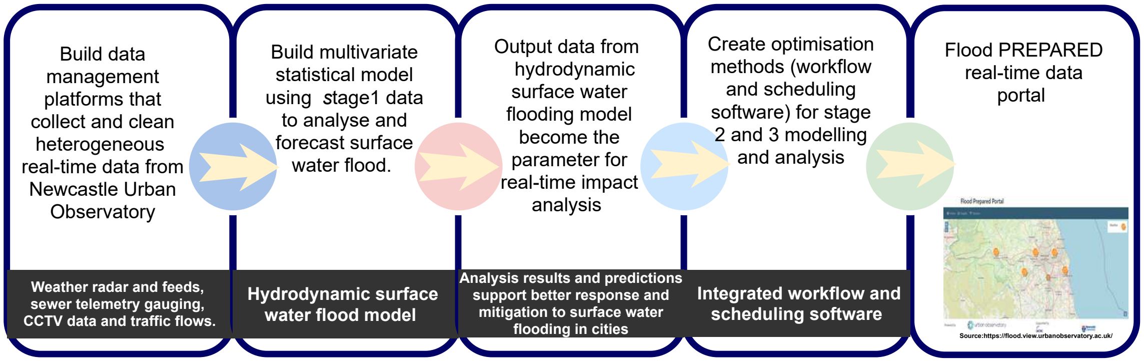

Build data management platforms that collect and clean heterogeneous real-time data from Newcastle Urban Observatory
Weather radar and feeds, sewer telemetry gauging, CCTV data and traffic flows.
Build multivariate statistical model using stage1 data to analyse and forecast surface water flood.
Hydrodynamic surface water flood model
Output data from hydrodynamic surface water flooding model become the parameter for real-time impact analysis
Analysis results and predictions support better response and mitigation to surface water flooding in cities
Create optimisation methods (workflow and scheduling software) for stage 2 and 3 modelling and analysis
Integrated workflow and scheduling software
Flood PREPARED real-time data portal
Flood Prepared Portal
Source:https://flood.view.urbanobservatory.ac.uk/

Figure 2: Predicting Rainfall Events by Physical Analytics of REaltime Data (Flood-**PREPARED) [56]**

archives a collection of historical datasets for many different metrics. These datasets acted as a reference for validating new predictions generated by James et al.'s dashboard. Overall, this dashboard aimed to multipurpose part of the observatory real-time data for crisis and disaster management [47, 57]. The same analyses were replicated across other cities such as Sheffield and showed similar outcomes. The Newcastle Observatory may provide aspects that can be replicated by observatories located in rural locations. First, a map for varied data and sensors that is interactive. Second, the capacity to download datasets in several formats. Third, the incorporation of live Twitter feeds.

3.1.2 Sheffield Urban Flows Observatory

Sponsored through the Engineering and Physical Sciences Research Council (EPSRC) and shares partnership with UKCRIC Universities [45, 37], Sheffield Urban Flows Observatory [58] actively sought to provide a carbon-free healthy environment. Hence, it developed a dynamic understanding of how flows of energy resources affect economic performance and social well-being. The observatory gathers, stores and analyses city data to monitor the interactive environmental performance of the city, engaging citizens and social systems. The technical platform captures real-time data that include air quality, weather, energy consumption, thermal and visual imaging. It consists of various types of sensors (fixed, mobile, and atmospheric), middleware (to gather, integrate data and transform them to meaningful information), data storage, and data analytics unit [58].

- Marvel, the sensing vehicle that measures the buildings heat signature and discovers their resource materials. The observatory management team and stakeholders can understand Sheffield's carbon footprint by linking Marvel's data with electric and gas demand information for various locations [58].
- Mobius, another mobile sensing vehicle that records radio frequencies, weather situations, and air quality.
- Weather sensing stations that assist decision-makers to identify local weather patterns that may influence air quality [58].
- Flying sensors or drones measuring Sheffield city air quality at different heights to locate and assess air pollution [58, 37].

Sheffield Urban Flows Observatory employed sensing vehicles and drones to monitor the city's environmental performance and collect data. The sensing vehicles traverse a variety of terrains, however, drones can reach inaccessible locations. Data collected from both tools complement each other to enable in-depth analysis. We propose that the use of drones and sensing vehicles in non-urban environments could be advantageous for animal conservation. Drones can be used to monitor animals for security concerns and deliver medications and other necessary supplies to remote workers. Similarly, sensing vehicles can collect observations like noise, position, and light data based on the sensors installed.

3.1.3 Bristol Urban Flows Observatory

UKCRIC Bristol Infrastructure Collaboratory [59] aims at transforming Bristol into a living laboratory that engages diverse communities from academia, businesses, and citizens. Using Open Data, Wireless Sensor Network (WSN), and smart technology solutions to address environmental and social sustainability concerns [59]. Data stream from various IoT networks adhere to the FIWARE [60] models. The system design relies on two main FIWARE components, (1) Context Broker (CB), which handles heterogeneous data and the multi-tenant users using the publish-subscribe approach. (2) IoT Agents construct IoT data internet protocols. Both components use NoSQL databases to store limited historical data. The Complex Event Processing (CEP) is federated with the CB to monitor data streams in real-time. CEP methods filter out the most relevant observations, detect interesting patterns and deduce relationships between events [42]. Few of Bristol Infrastructure Collaboratory's research initiatives include (i) a middleware to speed up deployment time in IoT Cloud platforms[9]. The middleware followed a bottom-up approach and achieved 50 percent improvement compared to the Unix-based bash shell scripting methods. The middleware consists of two nodeJS applications, Physical Resource Manager (PRM) and IoT Services Manager (ISM). Both applications operate from cloud servers. PRM allocates the required computation resources (e.g., RAM, CPU and storage) to the IoT application and ISM pairs the platform service IDs to the container created by PRM to isolate the platform's users. (ii) a live system to monitor the water quality of Bristol Floating Harbour, covering three locations[42]. Connecting to a WI-FI provided by Bristol Is Open (BIO) ICT infrastructure enabled researchers to combine the manual in-situ water monitoring methods with WSN to measure water quality at configurable frequency rates. Thus, achieved real-time data processing. Other related projects used the same data to create predictive water quality models to assist authorities in making evident decisions. (iii) a system to short-term monitor Clifton Suspension Bridge using the Structural Health Monitoring (SHM) [61, 62, 63]. SHM used APIs to combine wireless sensors and data management systems to collect, integrate, and display data about bridge loading usage. Stored data helped another project to predict the count of vehicles crossing the bridge [64]. The outcome platform is built up using Lord MSCL Python API [65], Apache Kafka [50], InfluxDB [66] time-series database and, Grafana Dashboard [67].

Further, the Walking On The Café Wall project [68] investigated the influence of the surrounding visual patterns on citizens' health and well-being. Researchers fitted a walkable corridor with "black and white" patterns and invited citizens to walk through it and provide feedback. Another project is the Residential damp detection system [69] where sensors network measured temperature and humidity in specific buildings. Incorporating sensors' readings provided accurate inferences of condensation that helped to decide the level of damp [59]. Conceivably, Bristol Infrastructure Collaboratory conducted a number of prototype initiatives near bespoke communication networks. In non-urban locations, the deployment of a WSN and the collecting of research data may be viable. For example, the data gathered from real-time monitoring of water quality can benefit both short- and long-term planning. Yet, network connections (e.g., WI-FI may not be available in the forest) and human resources (e.g., a shortage of field engineers and computer scientists) are potential roadblocks.

3.1.4 Cranfield Urban Observatory

Cranfield Urban Observatory [70] offers data-centric and remote sensing solutions for environmental, social, and economic matters. It has a well-established information technology unit to link many spatially distributed sensors. Its IoT network consists of different types of sensors fitted and connected to monitor noise and air pollution, water consumption and citizens observations. The observatory extracts the data from different sensors and publishes them in real-time along with dedicated analytics tools and visualisations. Domain experts can monitor the city environmental performance and make informed decisions to enhance life quality, health, and well-being [70, 71]. The observatory sponsored various projects and Cranfield University uses its data in teaching. For example, monitoring bats hunting time patterns, using ultrasonic acoustic sensors and machine learning algorithms [72], and measuring and comparing soil temperature during summer peak in multiple urban green spaces -using soil sensors and statistical analysis. Such projects fit well in the non-urban scenario. For instance, using real-time data acquired from acoustic sensors may help the bio-science research in creating, accurate predictive models.

3.1.5 Birmingham Urban Observatory

Birmingham has the UK's second-highest population after London. The city's high population density may put strain on infrastructure, public services, and the environment. As a result, borough administrators expend resources in controlling housing, transportation, health, and energy conditions in order to maintain sufficient living standards [73, 37]. In particular, Monitoring the environmental, economic, and social factors that may impact these critical infrastructures. The Birmingham Urban Observatory helps with this by serving as a tool that keeps track of a wide range of observable facts, like the weather, traffic flow, and biodiversity. The platform collects data from different sensors placed around the city and makes it available to the public through an interactive user interface. It also adds tools for analysis and a structure for governance to data. The observatory assists its users and stakeholders in exploring and analysing diverse data for the purpose of making informed decisions, engaging the public in information sharing, and initiating positive change. Observations made by sensors include air, soil, and grass surface temperatures, wind speed and direction, vapour pressure, sun radiation, and precipitation rate. Biodiversity sensing, including the detection of birds, can be gleaned from the Birmingham Urban Observatory and incorporated into non-urban data systems. It may aid researchers in accurately recording the migration behaviours of birds.

Table 3: Summarises and justifies recommendations to non-urban observatories, focusing on what can be learnt from urban observatory to apply on the non-urban settings.

Urban Observatory	Suggested features to non-urban observatories	Why
Newcastle [46]	Data ingestion, presentation and sharing	Inclusive and scalable (i.e., the system integrates different sensors and data sources and extendable to accommodate more features).
Sheffield [58]	Sensing vehicles and drones	Reach inaccessible locations.
Bristol [59]	WSN deployment and WI-FI outsourcing	To monitor water quality in real-time for making informed decisions.
Cranfield [71]	Monitoring animals using acoustic sensors	To use real data in teaching and wildlife research.
Birmingham [73, 78]	Real-time birds detector	To help in monitoring and analysing birds' migration and roosting patterns.
Manchester [74, 75]	Crowdsourcing and the Semantic Web approach	Engage citizens to have a say, integrate heterogeneous data sources and infer new events.

3.1.6 Manchester Urban Observatory

An interdisciplinary research hub [74, 75] that aims to collect, analyse and share urban data for decision support. Currently, the observatory runs a variety of themes in collaboration with other universities. The ongoing projects are:

- CF-Health-Hub, an electronic system that aimed to improve the health and well-being of patients diagnosed with Cystic Fibrosis (CF). The hub has been operating since 2015, with a mutual effort between six universities, twenty-three CF specialised centres and over a thousand patients. During the recent COVID-19 outbreak early 2020, it became critical to minimise face to face consultations and time in the hospital for CF patients. Therefore, the system sought to deliver virtual clinics with the help of the team in Manchester Urban University, which, in turn, integrated new medical tools into the hub platform [74].
- Evaluating mobile air quality measurements, the project intended to evaluate air quality across the city. It was motivated by the swift increase of air quality IoT devices with insufficient provenance data.
- Detecting biological particulates in the urban environment, this collaborative initiative is working towards applying a novel online technique for sampling biological particles.
- Quantifying citizen behavioural response to the city environment, an initiative to empower citizens to make responsible decisions about lifestyle choices. For example, encourage walking or cycling more than using vehicles.
- Health wearables, smart tools that volunteer patients or health practitioners can wear and collect readings. They support the investigation into instant health reactions from air pollution, achieved by linking readings captured from air quality sensors and the wearables.
- Monitoring well-being, citizens assess the level of public spaces usage across Manchester using cameras' data. The goal is to account for the health benefits of outdoor exercise and socialising.
- Air quality in Manchester's schools, the team at Manchester Urban Observatory are working together with 12 local schools to improve air quality and minimise children's exposure to traffic-related pollutants [76].

The dedicated observatory platform is known as "Manchester-I" [75]. It offers free and real-time air quality, flood monitoring, and traffic flow information. Manchester Urban Observatory was linked to Triangulum [34], a European Union-funded smart city data ecosystem. The Manchester Urban Observatory team has irreversibly rebuilt the platform and combined data from numerous sensors located throughout the city. They also created a web API that will leverage the semantic web technology's capability by using JSON-LD [77, 75]. The API provides its users with historical, real-time, and contextual data and assists them in discovering useful information about the data of interest. Manchester Observatory has two lessons to teach non-urban observatories. For starters, crowdsourcing is a low-cost strategy that allows citizens to directly influence problem solutions. Second, semantic web technology enables human and machine data understanding through interoperability and semantic data integration. To that aim, the semantic web's inference capabilities (e.g., if-then rules) allow users to enter rules to standardise data and deduce new events based on existing data.

3.2 Triangulum City Dashboard

European Smart Cities Communities Lighthouse Projects consist of fourteen members collaborating to develop Europe's future smart cities [34]. Triangulum is a member that entails three test-beds lighthouses located in Manchester (England), Stavanger (Norway) and Eindhoven (Netherlands). It brings together experts from the relevant fields with mutual interest to advance smart city plans and replicate them in fellow cities, Leipzig (Germany), Prague (Czechia) and Sabadell (Spain). The project focused on themes of energy, environment, and transport with an overall intention of improving citizens' life quality, using Open Data and technology [34]. More specifically, Triangulum intended to show that decisions based on real-time data and public engagement can save energy and lower cities' carbon dioxide (CO2) emissions levels. Mina et al. [34] introduced the Triangulum City Dashboard, the cloud platform that monitors Stavanger city and displays its real-time data to the public. The dashboard facilitates data acquisition and analysis through its complementary toolkit, enabling users to mine and explore diverse datasets from multiple sources. Data providers (public transport authorities and energy suppliers) send their sensors data to the dashboard cloud-based platform. The dashboard presents five different datasets from the transport and energy domains. Stavanger University researchers architected the platform as a three-layered, bottom-up structure to accommodate data streams for public transport, electric-assist cargo bikes, parking spaces, electricity consumptions, central energy plant and renewable power. It combined data integration, data quality, and data governance solutions that complied with the regulations of the General Data Protection Regulation (GDPR)[79]. Triangulum City Dashboard layers are described as follows:

- Perception Layer, contains the Triangulum data providers and their dedicated APIs [34].
- Processing Layer, crunches the data that Triangulum providers offer through their APIs. The data may arrive in different formats (e.g. CSV, JSON, XLXS); therefore, each format expects a corresponding configuration. Besides, if providers require specific system adaption to work with their data model, they must fill and submit a "Data Intake Form" shared in Google Docs. Accordingly, researchers at Stavanger University create tailored data solutions. Following adaption, data flow automatically from providers to the platform on a frequent and infrequent basis. Then, Logstash, a tool programmed on JRuby and can handle many types of data in the IoT, ingests the automated data, index, and store them in an Elasticsearch cluster. Elasticsearch is a search engine with an API and analysis toolkit that supports software multitenancy (multiple users share the same software) and NoSQL databases storage systems. Users can choose to access the datasets either through Elasticsearch or Kibana, the data visualisation console for Elasticsearch [34]. It manages data and their applications by performing central processes including data mining, filtering, sampling, analysis, and validation.
- Presentation Layer, the user interface graphically visualises data for decision support. It presents concerned domain data in a flexible and downloadable format, including tables, maps, and interactive charts.

The Triangulum project emphasises multi-stakeholder collaboration, diverse backgrounds, and support from the top down. These attributes could enhance the growth of non-urban observatories. In hostile circumstances where human and material resources may be scarce, such initiatives emerge. Governments, in particular, may be able to support the complete infrastructure (i.e., through funding research, network coverage, and IoT devices) despite the fact that different stakeholders bring distinct scientific and technological expertise).

3.3 Channel Coastal Observatory

The National Network of Regional Coastal Monitoring Programmes has existed since 2011 and fostered six active projects located across the English coastline. The collective goal is to collect in-situ coastal monitoring data. Contarinis et al.[80] stated that traditional management approaches exhibited some inconsistencies in the quality of the data collected and their methodologies. Channel Coastal Observatory [81], in turn, aimed to provide consistent and faithful data that can assist decision-makers in understanding the coastal behaviour and identify the potential risks of coastal flooding and erosion [81, 82]. Programme Coastal regions include the Northeast, East Riding of Yorkshire, Anglian, Southeast region (low-lying land), and Northwest. Data types collected and displayed on the Channel Coastal Observatory include topographic and hydrographic surveys. The former deals with beaches, cliffs, dunes, and coastal defence structures, while the latter expands from the Mean Low Water (MLW) contour to 1 Kilometre out sea [81].

To design the programme, managers created standard monitoring timetables and tailored them to each coastline nature and risks. Common nature comprises monitoring the coastal structure, geomorphology, while risks indicate exposure to wave attack and flood. They set up four management policies to classify coastal sites risk level- besides a different operational category named as a beach management plan, created for sites under certain agreements [81]. Even though coastlines differ in their local factors and managed risks according to their regions, yet the Channel Coastal Observatory included some hybrid monitoring approaches that can apply on most of the targeted coastlines [81]. Channel Coastal Observatory provides its end-users with a user-friendly interactive interface that unified access to real-time data and many other facilities, such as data catalogues, reports, and analysis toolkits.

Table 4: Channel Coastal Observatory programme composition that aimed to provide consistent and faithful data that can assist decision-makers in understanding the coastal behaviour and identify the potential risks of coastal flooding and erosion [81, 82].

Composition	Purpose
Beach profile	Beach profile contains measurements of slope along a cross-shore transect. Combing beach profiles with other topographic data reduced records duplication and assisted a smooth data integration in a single platform.
Waves monitoring	Channel Coastal Observatory publishes real-time wave data. Government institutions use them for many purposes, including forecasting environmental hazards such as flood warnings.
Tides monitoring	Coastal monitoring programmes installed A-class tides gauge network to capture real-time tide data in limited locations and under careful formal consideration. Due to regulations constraints, tide gauges were deployed at Port Issac and provided the only real-time coastal tidal data between Land's End and Ilfracombe. Channel Coastal Observatory transmits them immediately to the International Oceanographic Commission's Sea Level Monitoring Facility to supplement tsunami warning data services.
Lidar	Lidar is useful for high-risk surveying areas where physical access is unsafe. Furthermore, Lidar can detect soft and swiftly eroding cliffs.
Aerial images	Regular aerial surveys help in evaluating cliff frontage erosion. Aerial images linked with Lidar can substitute topographic data.
Satellite imagery	Useful for monitoring moving sandbanks at Morecambe Bay.
ARGUS cameras	Useful for monitoring the seawall scour at Cleveleys.
Laser scanners	Topographic surveys rely heavily on laser scanners to acquire high-density beach data such as cliffs structures.

Table 5: Linked Data Finland [85] proposed star model that built on Tim Berners-Lee [19] and explained at section 2.

Stars	Linked Open Data
⋆⋆⋆⋆⋆⋆	Hyvönen et al. [85] used the Live OWL Documentation Environment (LODE) to document the Linked Open Data with their tailored schemas. The associated schemas come as HTML files and contain lists of classes, properties, and axioms, obtained from OWL ontologies.
⋆⋆⋆⋆⋆⋆⋆	Hyvönen et al. [85] created a web-page to validate Linked Open Data against the schemas and analyses their documentation.

3.4 Grow Observatory

Grow Observatory [83] is a citizens science project that secured funds from the European Commission. It aims to build a citizen observatory system for measuring in-situ soil moisture [84]. Citizens observatory refers to stakeholders (e.g. civilians, scientists, and policy-makers) working together on a research. Grow Observatory priorities the engagement of a broad range of users and raises awareness about the advantages of environmental monitoring. It extracts in-situ data for satellite validation and creates a mobile app for real-time interaction. Stakeholders employ soil sensors to collect data, store them in databases, access them through a mobile app and data portals. Grow Observatory developed two soil sensing network approaches, the Flower Power and the Do-It-Yourself (DIY) sensors. The observatory team fit the Flower Power sensors in the soil dirt to record the moisture, light, and air temperature every quarter-hour. Each sensor has eighty days of data capacity that users can access remotely through a mobile app with a Bluetooth connection [84]. Do-It-Yourself (DIY) sensors implied that stakeholders install commercial sensors in decided locations and manage their data independently. They store the data into the Grow Observatory platform for further integration and analysis. Grow Observatory managed to integrate the various sensors data allowing its members unified access via an online data platform. The members can register their sensors, store data, access and download them via the *MyData* download tool. Then, Grow sensor database requests and stores these data. The collaboration hub obtains the Grow sensor database data and presents them in the members' separate web pages. The visualisations available are the time-series, graphs of the sensor's measurements and their location on the map. Furthermore, an edible plant database provides the observatory mobile app recommendations about plants locations and seasons, while a Land survey database accommodates surveyors' data from their mobile app. *GEOSS*, the observatory's dedicated data portal, sequentially, provide public access to the Grow Observatory archived earth observations data.

3.5 Cyclone Dashboards

Tilley et al. [36] created Cyclone Dashboards to monitor the major centres across Australia and the Pacific Islands in response to the Tropical Cyclone Debbie strike on the Coral Sea, northern Australia coast in 2017. The extreme event claimed fourteen lives and caused significant damage to resources and properties. Bureau of Meteorology Agency announced its category to be four storms. Here, the authors argued that citizens had no unified access to various real-time information and Open Data sources before and during the disaster. In the presence of such tools, decision-makers may have predicted the cyclone early signs by linking, for instance, weather data with wind speed and citizens real-time feeds. Further, communication with the public could have been quicker to warn them against the hazard. Hence, Tilley et al. [36] built the Cyclone Dashboards under the wire and in the last possible moment to be prepared for any future waves. Cyclone Dashboards source their data from several providers, including the Bureau of Meteorology, Twitter, and Google website. Bureau of Meteorology offered cyclone tracking map and advice, rain radar, wind, weather, and tides forecast while Twitter and Google supplied information about traffic advice and condition, respectively. The dashboards aggregate the heterogeneous data at the Extract-Transform-Load (ETL) [36] layer. LAMP (Linux, Apache, MySQL, and PHP) stack manages the incoming diverse and real-time data and presents them publicly on a single screen [36, 86].

3.6 Linked Data Finland

Linked Data Finland [87] is a research data service platform built with the semantic web technologies. In an enduring collaboration between universities and businesses, the platform aims to support the publishers and consumers of structured data. Hyvönen et al. [85] stated that Linked Data re-users often face two main obstacles, these are, understanding the complexity of datasets characteristics and assessing the sufficiency of data quality for the intended purposes. For this reason, Hyvönen et al. added two more stars to the 5-star data model by Tim Berners-Lee [19] explained at Section 2. To earn the 6th star, authors documented the open datasets with programed schemas that explicitly describe their variables using the Live OWL Documentation Environment (LODE) [85], an open-source paradigm that automates the extraction of classes, properties and axioms from OWL ontologies and represents them in Hypertext Markup Language (HTML) files. The 7th star validated data of interest against the programed schemas by setting up a homepage that analyses the Linked Data documentation and reports the vocabulary usage.

The outcome product was a data portal named Linked Data Finland [87]. The portal automates the process of publishing different topics of linked and opened datasets about Finnish history, law, science, ornithological observations, weather, and news. The datasets come with their associated metadata and data curation tools. The system core processes explain as follows:

- Data publishing automation, the portal receives datasets from publishers in an RDF format with limited metadata. Then, it applies the latest versions of W3C Service Description recommendation, and Vocabulary of Interlinked Datasets (VoID) [88, 89] stores them in triple store databases, accessed and queried via a SPARQL endpoint. Alternatively, the portal generates datasets and graph names list in JSON, containing data labels and descriptions. Linked Data Finland also has a webpage to each dataset offering links to downloadable sub-datasets and visualisations (license condition may apply), links to available data schemas, documentation, and reports datasets' inspection in various RDF serialisations forms [90] (e.g., Turtle, RDF/XML, RDF/JSON, N3, N-triples).
- Data Curation, denotes the creating, managing, and validating tasks performed on data of interest. Linked Data Finland portal encompasses many tools for managing data and generating semantic annotations. For example, Seco Lexical Analysis Services20 for natural language processing and SAHA22 for real-time interaction, authors in [85] altered SAHA to become a Linked Data Browser in the portal.

3.7 Linked Open Data-Based Web Portal for Sharing IoT Health and Fitness Datasets

Reda et al. [91] created an online data portal using the semantic web technologies and Linked Data, referencing the IoT Fitness Ontology (IFO) [92, 93]. The portal aims at integrating heterogeneous IoT big datasets and sharing them freely with communities of researchers and decision-makers in a structured format [91].

The web portal consists of the following modular layers where each layer is upgradeable or changeable separately without affecting the entire system [91].

- Perception Layer, gathers IoT health and fitness data from citizens manual input or cloud servers encoded retrieval systems.

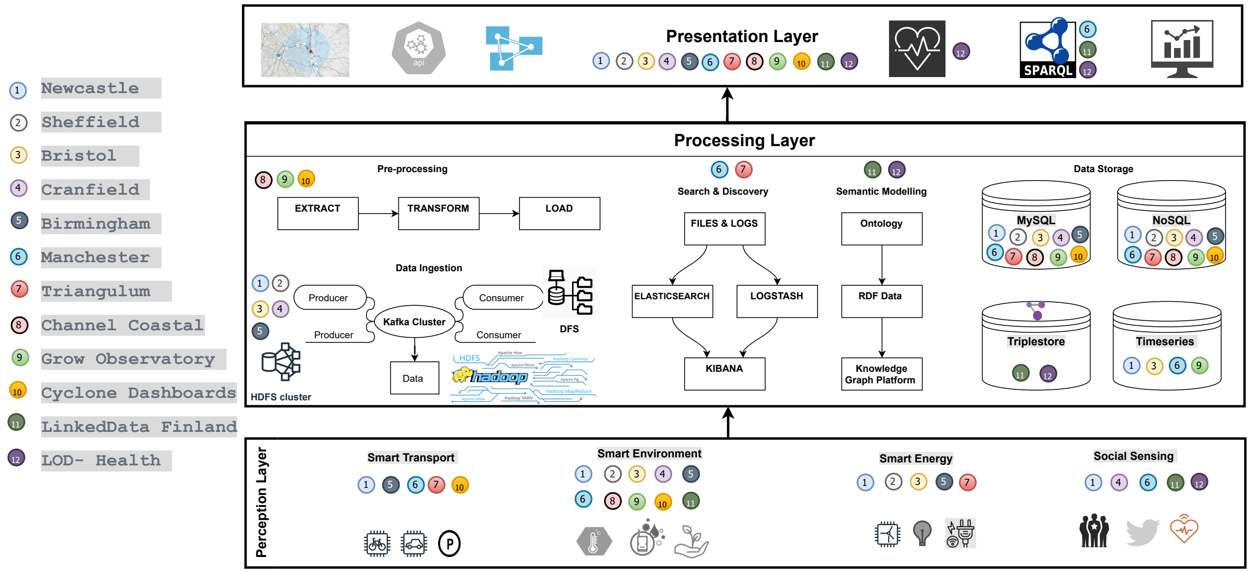


Figure 3: Open Data Observatories' data management conceptual design. In this bottom-up system schematic, we have condensed the twelve reviewed observatories. We identified them by assigning each observatory a unique symbol (i.e., numerals within a circular frame) in no particular order. These symbols then represented the data source, processing, storage, and presentation strategies used by each observatory. For example, we assigned the symbol 11 in a circle to Linked Data Finland, which was placed at the IoT Big Data, Semantic Web, and SPARQL endpoints at the perception, processing, and presentation layers, respectively.

Table 6: Summarises and justifies recommendations to non-urban observatories, focusing on what can be learnt from these data observatory to apply on the non-urban settings.

Data Observatory	Suggested features to replicate in non-urban observatories	Why
Triangulum [34]	Multi-stakeholder partnerships	Encourage global expansion.
Channel Coastal [81]	Extreme events analysis	Useful for natural disaster management
Grow [84][83]	Do-it-Yourself (DIY) sensors deployment	Encourage citizens engagements.
LOD Finland [85]	Data documentations and validation tools	Maintain fit data provenance and quality
LOD IoT Health [91]	Semantic web processing	Ability to link different kinds of data
Cyclone [36]	Real-time data presentation	Useful for predicting natural disasters

- Processing Layer, accommodates the mapping function where raw semi-structured data transform into RDF graphs with semantic annotation from the IFO. Then get archived in the designated databases (RDF triple store) and queried via a SPARQL endpoint.
- Presentation Layer: enables intended users to query and visualise the stored RDF data via the purpose-built dashboard.

The following section explores the data integrated by our reviewed observatories. We investigate and compare their data types and insights for the various applied domains sources, transport, energy, environment, and social sensing.

4 Data Domains and Insights

Many stakeholders, including consumers, governments, and academia, create and share an Open Data Observatory [11]. One of the main components for acquiring data in an observatory are the wireless sensors embedded in the smart devices [94]. Any stakeholder may install sensors or smart devices for specific reasons, and other stakeholders' platforms and APIs may manage and use them for different purposes. That said, all involved participants are likely to use their data platforms. These smart devices measure various metrics for many domains, including transport, environment and energy. Figure 4 visualises four common domains with their data types covered by the reviewed Open Data Observatories.

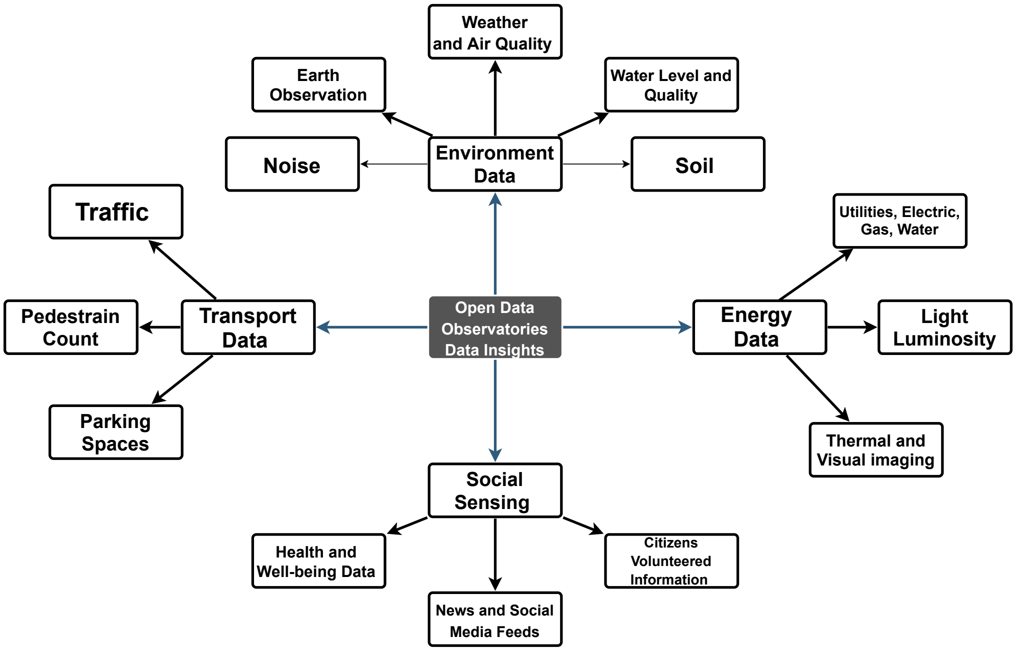


Figure 4: Some of data types in four applied domains at Open Data Observatories

4.1 Transport Data

Transport involves any activities that take place outside our homes locally or abroad. For example, daily commutes to work whether on foot, cycling or using vehicles- and travelling abroad by cars, ships, jets, and trains. Transport data exist in a quantifiable manner, are often re-used, and have a significant impact on our daily life [95]. Open Data Observatories databases obtain transport data, and metadata from diverse sources include sensor platforms and citizens volunteering information. Transport data entail traffic flow, vehicles count, public transportation, parking spaces, congestion, average speed, journey time and pedestrian count. Table 7 signifies transportation data type collected by some of our reviewed observatories. IoT sensor nodes network often consists of expensive and economical sensors from several suppliers. These sensors cover specific geographical areas, connected wirelessly to capture and record real-time observations. Newcastle Urban Observatory, for instance, collects transport data - sourced from their deployed sensors throughout North East of England and the sensors of NE Travel Data API - and publish them at the observatory website. Large-scale heterogeneous transport data arrive in the observatory with different metrics and high update frequency. Citizens can access and download real-time data to help them make smart decisions such as planning a journey with the shortest path, avoiding congestion, and finding local parking spaces. A further example of the advantages of transport data is operational at Triangulum city dashboard, its smart parking network analyses five sensors data for differently located car parks in Eindhoven. Then, publish a real-time dataset containing the car parks names, UNIX time, capacities, geographic coordinates (longitude-latitude) and available spaces. The dashboard visualises these variables to road users via its integrated map [34].

4.2 Social Sensing Data

Social sensing is mainly about engaging citizens and collecting data from them. This process can be achieved voluntarily or involuntarily through digital services (e.g. social media platforms, emails, electronic forms) and manually via paper questionnaires and surveys. IoT has played a significant role in acquiring social sensing data, the micro-sensors embedded into smart devices electronic boards (e.g. mobile phones, IPADs) record and communicate user's data remotely. For example, some global positioning system (GPS) enables mobile phones to capture satellite signals to track citizens geographical location and movements – uses motion sensors for broader area coverage and more accurate positioning. A recently developed dashboard [57] derived from Newcastle Urban Observatory data purposed to quantify the impacts of COVID-19 measures imposed by the UK government. The system relies on home sensors, machine learning algorithms and artificial intelligence to detect people movements in Newcastle city. The dashboard monitors local streets and captures citizens mobility, allowing decision-makers real-time insights. For example, they can monitor if citizens are adhering to COVID-19 restrictions such as social distancing [57, 96]. Social media platforms gather huge data about their users, starting from the registration to the posts. For instance, Twitter posts can under go sentiment

Table 7: Transport data types at Open Data Observatories

Open Data Observatory	Traffic	Pedestrian Count	Parking Spaces
Newcastle Urban Observatory [46]	✓	✓	✓
Bristol Urban Flows Observatory [59]	✓		
Birmingham Urban Observatory [73]	✓	✓	
Manchester Urban Observatory [74] and Manchester-I [75]	✓		
Triangulum City Dashboard [34]	✓		✓
Cyclone Dashboards [36]	✓		

Table 8: Social sensing data types at Open Data Observatories

Open Data Observatory	Health Well-being	Social Media	Citizens Data
Newcastle Urban Observatory [46]	✓	✓	✓
Cranfield Urban Observatory[71]			✓
Linked Data Finland [85]			✓
LOD for IoT Health and Fitness [91]	✓		✓
Cyclone Dashboards [36]		✓	✓

analysis to estimate whether the contents are negative, neutral or positive [97]. Collaborative Online Social Media Observatory (COSMOS) [98] collects, analyses and presents social media data to address future research questions. Table 8 lists some social sensing data associated with our reviewed observatories.

4.3 Environment Data

At Open Data Observatories, environment data attract the attention of researchers, stakeholders and end-users. In particular, the environmental monitoring, climate change [86] and their connection to other data domains (e.g. transport and health). For example, certain types of vehicle fuel may pollute air causing low air quality that impacts ventilation and circulation, thus may negatively affect health. Also, weather conditions may affect traffic flow and people movements. Environment data involve climate conditions, earth observation, air quality, water levels, soil moisture, and organisms' activities. Some observatories collect and combine multiple environmental data to predict extreme events, such as floods, wildfires, and severe storms, and protect living creatures' welfare. Consequently, taking actions towards preventing them or minimizing their impact on habitat. For example, Taneja et al. [99] built a fog based IoT platform that collect data from collars on cows to monitor their health and wellbeing. Cyclone Dashboards [36] integrated multiple Open Data about the weather, wind, tides, and rainfall to assist in predicting severe cyclones in the Australian coasts. Grow Observatory [84, 41] engaged citizens in disparate locations to volunteer the collection of in situ soil conditions data. Newcastle Urban Observatory integrated and published diverse environmental data, attached with their metadata for decision support, and equipped researchers, businesses, and the public with free real-time urban data. The observatory deployed a large-scale network of over 500 IoT sensors recording observations ranging from weather, air quality to pedestrians' movements [95]. Figure 5 shows the environment data types, and parameter counts at Newcastle Urban Observatory. Furthermore, table 6 lists examples of the data types' parameters and their measuring units. Raw data were obtained from (https://newcastle.urbanobservatory.ac.uk/api-docs/doc/sensors-dash-types-csv/).

4.4 Energy Data

Energy data gathered by Open Data Observatories include electricity and gas consumption. Observing these factors helps to identify areas that may require more attention. For example, the Triangulum City Dashboard [34] fitted smart energy meters across Stavanger city that record usage every 10 seconds—then analysed over a year worth of data collected from 56 residences, exploring patterns trends. Nevertheless, the UKCRIC observatories in Sheffield [37] and Bristol [59] monitor and record energy usage, thermal, visual, and hyperspectral mapping, while Birmingham [73] focuses on sensing light luminosity. Another example is the scalable energy data platform developed by Zhang, Y.-Y. et al. [100]. It senses, integrates and shares isolated and heterogeneous energy consumption data from smart buildings. Here, sensor nodes collect and communicate real-time data, NoSQL database stores important information using a unified data schema. This framework enables real-time interaction with add-on web services tools to support data mining (i.e. extracting patterns from multiple sensors at multi-scale data) and analysis (i.e. customised reporting that may include statistical analysis and forecasting). Table 10 lists examples of energy data types at the Open Data Observatories. The following section outlines data management at Open Data Observatories- identifying their data sources, formats, and designated databases. We also reviewed the data processing methods along with the analysis and visualisation.

Table 9: Comparison of Data collected at the Open Data Observatories

Data Observatory	Transport Data	Environment Data	Energy Data	Social Sensing
Newcastle [46]	**Traffic** (vehicles count, parking spaces, traffic flow, congestion, average speed, journey time); **Pedestrian Count** (people geographic walking directions);	**Weather** (temperature, rainfall, rain accumulation, visibility, wind speed and direction, humidity, dew point, sunshine hours, solar radiation, pressure); **Air quality** (e.g. PM 4, NO2); **Water quality** (e.g. temperature, depth, conductance, dissolved oxygen); **Seismic** (e.g. horizontal and vertical displacement); **Sewage levels**; **Soil** (e.g. soil moisture, temperature, CO2 range); **Noise** (sound); **Water Level** (river, tidal level); **Beehives**.	**Electricity** (real power); **Buildings** (utilities).	**Social media feeds**; **Employees feedback** (health and well-being in office environment); **Quantifying the impacts of CORONAVIRUS (COVID-19) measures**.
Sheffield [58]		**Air quality**; **Weather** (local weather conditions).	**Thermal and visual imaging**; **Energy usage**.	
Bristol [59]	**Traffic flow**	**Air quality**; **Weather**; **Lidar**.	Thermal, visual, and hyper-spectral mapping.	
Cranfield [71]		**Water usage**; **Air and noise pollution**; **Soil moisture**.		Customers satisfaction.
Brimingham [73]	**Traffic** (vehicle count); **Pedestrian Count**;	**Air pollution**; **Heatwaves**; **Flood monitoring**; **Weather Station**; **Rainfall radar system**; **Lighting Detection**; **Rail moisture sensing**; **Acoustic underground sensing system**; **Automatic Passive Integrated Transponder (PIT)** tags readers; Birds Detectors	**Light luminosity**; **Lux meters** (light meters).	
Manchester [74, 75]	**Traffic flow**	**Air quality**; **Weather**; **Flood monitoring**.		Crowdsourcing.
Triangulum [34]	**Public transport**; **Parking management**; **Carbon emissions**;**Usage of electric vehicles and charging infrastructure** (e-bikes, e-buses, e-cars).		**Carbon emissions**; **Renewables** (heating, cooling, and electricity).	
Channel Coastal[81]		Ortho-rectified aerial and False Colour Infra-red imagery , Non-rectified aerial imagery, Oblique imagery, Bathymetry data, Photogrammetric profile data, Topographic survey data, Waves, Tides, Meteorological data Wave data; **Lidar data**; **Topographic and hydrographic surveys**		
Grow[84, 83]		**In-situ soil moisture**;**Air temperature**; **Soil fertilizer level**.		
Linked Data Finland [85]		**Ornithological** (bird-watching); **Weather**; **Science Linked Open Data**.		**Linked news**; **History and law Linked Open Data**.
Linked Data Health and Fitness [91]				Measurements of body weight, blood pressure and heart rates.
Cyclones Dashboards [36]	**Traffic** routes and conditions	**Rain Radar**; **Wind**; **Weather**; **Cyclone** tracking map and advice and **Tides Forecast** .		

Table 10: Energy data types at Open Data Observatories

Open Data Observatory	Utilities	Light Luminosity	Imaging
Newcastle Urban Observatory [46]	✓	✓	✓
Sheffield Urabn Flows Observatory [58]	✓		✓
Bristol Urban Flows Observatory [59]	✓		✓
Birmingham Urban Observatory [73]	✓	✓	
Manchester Urban Observatory [74] [75]	✓		
Triangulum City Dashboard [34]	✓		

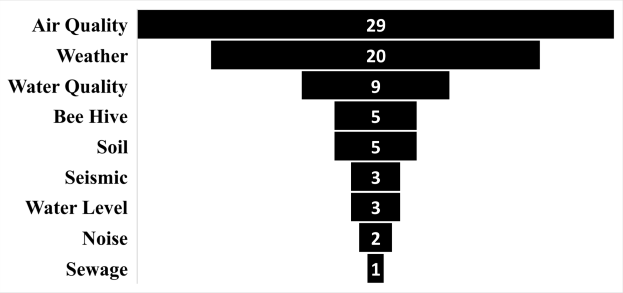
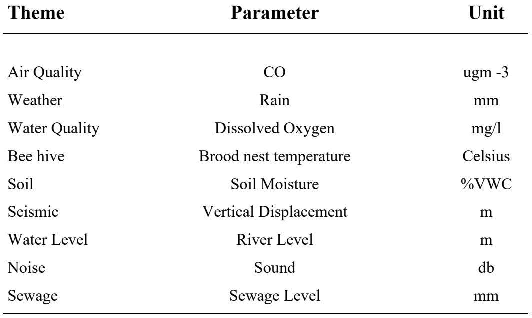


Figure 5: Newcastle Urban Observatory parameters count by theme [46]



Theme	Parameter	Unit
Air Quality	CO	ugm -3
Weather	Rain	mm
Water Quality	Dissolved Oxygen	mg/l
Bee hive	Brood nest temperature	Celsius
Soil	Soil Moisture	%VWC
Seismic	Vertical Displacement	m
Water Level	River Level	m
Noise	Sound	db
Sewage	Sewage Level	mm

Figure 6: Newcastle Urban Observatory parameters examples and their measuring unit [46]

5 Data Management in Open Data Observatories

Open Data Observatories represent standardised bottom-up systems. Their heterogeneous data escalate from multiple sources from the perception to the processing layer, which connects and operates various devices, resources, and systems [101]. The processing layer fosters several functional requirements during data collection, aggregation and storage. Namely, resource discovery management at the collection process [102], the management of data, event and code at the aggregation and storage stage [101]. Primary categories for standards protocols for discovering and configuring IoT devices are explained in [103]. IoT sensors transmit their collected data to different locations, while inside the same sensors network, they use communication protocols such as the IEEE 802.15.4 standard, Zigbee [104, 105, 103]. This section covers the data management elements from the data sources, generated data formats, and databases. Further, it discusses the various processing approaches applied by the reviewed observatories besides the predictive analysis and visualisation techniques. Nevertheless, it classified data management approaches in taxonomy 8, to assist data observatories' stakeholders in exploring available options.

5.1 Data Sources

Open Data Observatories source their data from open data portals, wireless sensor networks and smart devices. Wireless Sensor Networks (WSNs) play a crucial role in data collection for the environment, transport and energy domains [106]. For instance, the 3600 sensors in Newcastle Urban Observatory [46] - measuring different physical environments stream various types of data. Manchester [74] has a variety of smart devices, including but not limited to - the microAeth AE51 with fitted sensors to monitor aerosol Black Carbon concentration in real-time and the ARISense that measures multiple climate pollutants such as nitrogen dioxide (NO2), and carbon monoxide (CO). At the Grow Observatory [41, 84], the Flower Power sensors measure in-situ soil moisture, level of fertiliser and air temperature every 15 minutes. Other technologies contributing data to such observatories include Lidar, ARGUS cameras and satellites. In addition to social media platforms (e.g.Twitter and Facebook) and citizens digital and paper reporting systems. Table 12 lists and compares the observatories data sources.

5.2 Data Formats

Open Data Observatories typically handle digital data in diverse formats to deliver innovative data services, visibility and transparency. In this context, each observatory's data formats depend heavily on its primary purpose. For example, Newcastle Urban Observatory [46] provides data repositories in machine-processable formats such as JSON and CSV. Linked Data Finland [85] serves completely different formats, that is, the Resource Description Framework (RDF) and Linked Open Data (LOD). Triangulum [34], and Cyclone Dashboards [36] focus on data visualisations that conform with their urban policies, so the underlying data may not be available in machine-processable formats. Noteworthy, data formats may require transformation when collected from their sources. For example, Triangulum city dashboard data created tailored adapters to ingest the different data formats [34]. Many observatories owners [46, 84, 81, 57, 36, 93] also applied scientific methods such as statistical analysis and machine learning to curate, transform and re-purpose the captured data. Curated data are then stored in designated databases according to their format. In what follows, we explain the databases types and illustrate how Open Data Observatories store their data depending on their formats.

Table 11: Lists and compares the Open Data Observatories data sources. Newcastle, Manchester and Cyclones observatories have the largest varieties of data sources having open Data and WSNs led the primary sources of data in the reviewed observatories.

Open Data Observatories	Reference	Open Data	Wireless Sensors	Smart Devices	Citizen Data	Weather Stations	Smart Cameras	RFID	Satellite/Lidar	Social Media	Sensing Vehicles	Drones	Crowdsourcing
Newcastle Urban Observatory	[46]	*	*	*	*	*	*	*	*	*			
Cyclone Dashboards	[36]	*	*	*	*	*	*	*	*	*			
Manchester UO and Manchester-I	[74] [75]	*	*	*	*	*	*	*	*				*
Channel Coastal Observatory	[81]	*	*	*	*	*	*	*	*				
Sheffield Urban Flows Observatory	[58]	*	*	*		*	*	*			*	*	
Grow Observatory	[84][83]	*	*	*	*	*	*		*				
Bristol Urban Flows Observatory	[59]	*	*	*	*	*				*			
Cranfield Urban Observatory	[71]	*	*		*	*							
Birmingham Urban Observatory	[73]	*	*	*		*	*						*
Triangulum City Dashboard	[34]	*	*	*	*		*						
Linked Data Finland	[85]	*			*								
Linked Open Data for IoT Health and Fitness	[91]	*			*								

Table 12: Lists and compares each Open Data Observatories data formats and designated databases

	Data Formats									Storage Databases				
Open Data Observatories	CSV	XIXS	JSON	JSON-LD	RDF	TEXT	Multimedia	Shapefiles	Time-series	MySQL	NoSQL	In-memory	Triple-stores	Time-series DB
Newcastle Urban Observatory[46]	*	*	*			*	*	*	*	*	*	*		*
Sheffield Urban Flows Observatory [58]	*	*	*			*	*	*	*	*	*	*		*
Bristol Urban Flows Observatory [59]	*	*	*			*	*	*	*	*	*	*		*
Cranfield Urban Observatory[71]	*	*	*			*	*	*	*	*	*	*		*
Birmingham Urban Observatory [73]	*	*	*			*	*	*	*	*	*	*		*
Manchester UO and Manchester-I [74] [75]	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Triangulum City Dashboard[34]	*	*	*			*	*	*	*	*	*	*		*
Channel Coastal Observatory [81]	*					*	*	*	*	*	*			*
Grow Observatory [84][83]	*							*	*	*	*	*		*
Linked Data Finland [85]					*								*	
Linked Open Data for IoT Health and Fitness[91]					*				*			*	*	*
Cyclone Dashboards[36]	*		*			*	*	*	*	*	*	*		

5.3 Storage Databases

Storage databases refer to organised data spaces with user interfaces that store specific data formats compatible with their design. In general, databases - according to their types- accommodate structured, semi-structured and unstructured data. They interactively arrange data to enable their authorised users to access, modify and update records. Databases widely exist in systems that process transactions, i.e. people records, and data warehouses that store integrated datasets for analysis and modelling [107]. For example, relational database management systems (RDBMS) [108, 16] launched in 1970, can only work well the structured data, such as XLXS, CSV and JSON, unlike the modern non-relational NoSQL (Not Only Structured Query Language) [108, 17], which emerged mid-2000 and can handle a tidal wave of data in all forms. The devices that generate data influence the storing and processing method through their data format. For instance, In [62], time-series data produced by sensors have to be stored in time-series databases such as influxDB for further processing and analysis. Grow Observatory [41] relied on in-memory sensor storage and remotely hired cloud servers [49] to store real-time and historical data, respectively. Researchers such as Jiang et al. [10] implemented a hybrid IoT data storage framework that integrates structured and unstructured data. Table 12 lists various databases and the corresponding format used by our surveyed observatories. As seen in this survey, the Open Data Observatories used query-based databases to fulfil applications request. For observatories that provide real-time data, self-querying databases would be more aware of real-time events and make smart decisions (e.g, detect extreme values and flag them).

5.4 Data Processing

Most Open Data Observatories execute their data in cloud platforms and rely heavily on edge computing [49, 41, 71] for real-time processing. Processed data are sent to the cloud through fog computing. Fog computing is a middle

Table 13: A comparative list for the Open Data Observatories processing and modelling approaches.

Open Data Observatories	References	Complex Event	Semantic Web	Context-Aware	Stream Processing	Time-series Analysis	Machine Learning	Optimisation	Visualisation	Statistical Analysis	Extreme Events	Simulation
Newcastle Urban Observatory	[46]	*		*	*	*	*	*	*	*	*	*
Sheffield Urban Flows Observatory	[58]	*		*	*	*	*	*	*	*		
Bristol Urban Flows Observatory	[59]	*		*	*	*	*	*	*	*		
Cranfield Urban Observatory	[71]				*	*	*	*	*	*		
Birmingham Urban Observatory	[73]	*		*	*	*	*	*	*	*		*
Manchester UO and Manchester-I	[74] [75]	*	*	*	*	*	*	*	*	*		
Triangulum City Dashboard	[34]			*	*	*	*	*	*	*	*	
Channel Coastal Observatory	[81]					*	*	*	*	*	*	
Grow Observatory	[84][83]				*	*	*	*	*	*	*	
Linked Data Finland	[85]		*						*			
Linked Open Data for IoT Health and Fitness	[91]		*						*			
Cyclone Dashboards	[36]	*		*	*		*		*	*	*	

layer between the edge and the cloud. It examines and filters out the relevant data to be transmitted to the cloud. The irrelevant data are either wiped out or analysed at the fog as their final destination [109]. Historical data processing methods occur in the backends of the cloud platforms. They manage data streams, schedule and automate tasks. Inversely, the graphic designs, query and search engines take place in the frontends. In other words, frontends enable end-users to interact with such platforms and obtain reusable data. They may also provide interactive visualisations and analysis toolkits. Each of our reviewed Open Data Observatories dealt with data processing dissimilarly. For example, Newcastle Urban Observatory [46] used parallel and distributed systems applications such as Apache Kafka [50], Hadoop [110, 111, 112, 49], to integrate and process their big, fast arriving, and heterogeneous data. Grow observatory[84, 83] and Channel coastal Observatory operates on hired servers and outsources their services. Grow Observatory [84, 113], for instance, partakes structure from the Geo-wiki.org [114]. Triangulum City Dashboard [34] ingested their diverse data with the help of Logstash, Elasticsearch, and Kibana- (ELK) Stack [115]. Hyvönen et al. [85] and Reda et al.[91] built their online data platforms leveraging Linked Data and the semantic web technologies to achieve interoperability, data conceptualisation, and linkage with other web data on a global scale [101]. Our reviewed observatories used a wide range of various processing techniques to suit their data. Interestingly, some of them such as [85] and [91, 75] shared the same processing methods (e.g., semantic web). Noticeably, Manchester Urban Observatory [74], and its data portal Manchester-I [75] are using all reviewed data formats and databases- this could be deduced from the recent expansion after separating from the Triangulum project and the adoption of the semantic web technologies. Table 13 demonstrates the processing and modelling approaches by the Open Data Observatories.

5.5 Predictive Analytics

In most Open Data Observatories, complex event processing analyses heterogeneous data fast enough to find interesting patterns [116]. However, information that can aid decision-making or inferring future events requires historical data. Predictive analytics play a crucial role in exploring data, recognising interesting patterns, and generating predictions. For example, the flood-prepared scheme by the Urban Observatory applied statistical and optimisation methods to build predictive models from heterogeneous real-time data feeds [56]. Nevertheless, visualised the model via an interactive dashboard that contains a data map. The curated data published on Manchester Urban Observatory rely on statistical analysis and machine learning to enable users to explore entities and their correlations, visualise their time-series and even choose to have missing values imputed via interpolation [117]. Bristol Urban Flows Observatory uses its various sensors real-time and historical data to monitor water quality [42] and predict the number of vehicles crossing Clifton Suspension Bridge respectively [59, 61]. Channel Coastal Observatory [81] analyse extreme values in its time series to predict tides. Relatively, Cyclone dashboards [36] aim to predict extreme events-drawing lessons from the past cyclone, Debbie, in 2017; these dashboards seek to communicate early warnings of future natural hazards via a unified display.

5.6 Visualisation

Data visualisations transform information into meaningful graphical representations that intended audiences can interpret [118]. Visualisations include static and interactive maps [119], charts such as time series, scatter plots, histograms [120], bar, and pie graphs. They can be performed by numerous amount of software packages ranging from Microsoft Excel, Matlab, SPSS [121] to the programming languages like Python and R. A good visualisation can be more descriptive

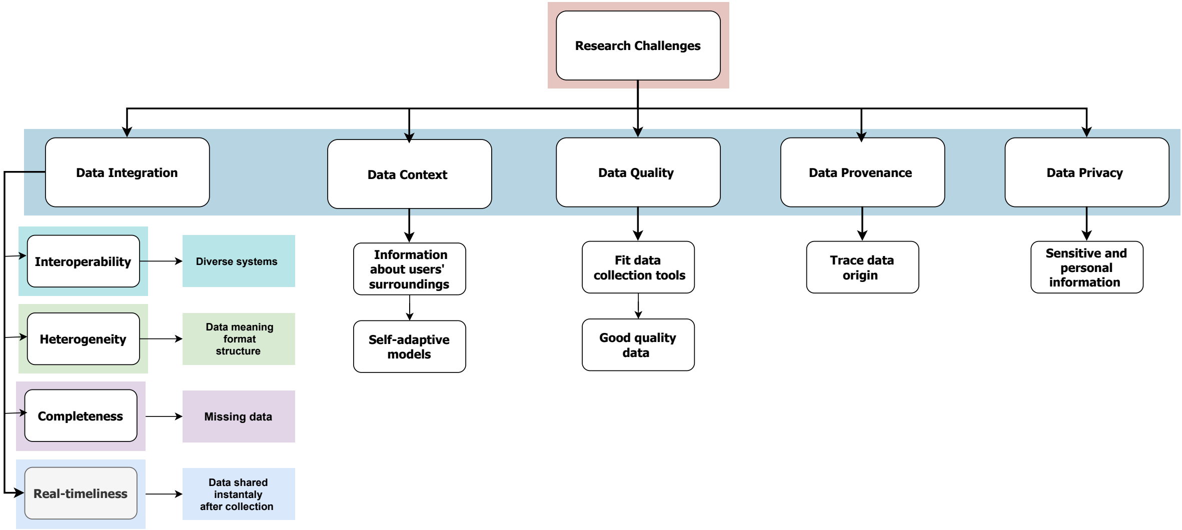


Figure 7: Research Challenges examples in the Open Data Observatories

than a text in communicating real-time events and research findings. For example, it may show correlations between variables, an up or down trend, a repeated pattern over time, normal distribution, right or left skewness and outliers. Furthermore, interactive visualisations can respond to the user manual and automated data updates and plot them against past observations for comparisons [122]. To design visualisations that fit the dynamic nature of data generated in observatories [5], deemed to be a challenging task due to the data large volume, heterogeneity, and high dimensionality [123, 124]. Examples of research efforts in Newcastle Urban Observatory [46] website. It contains an interactive map for different sensors - updates in real-time, a live Twitter feed and active links to the datasets, radar, dynamic time series and many more. Grow Observatory [41], which not only developed dynamic maps and visualisations for its stakeholders but also shared its standardised sensors data to promote interoperability of Grow data and other databases such as UK Met. Reda et al. [91] visualised their IoT health and fitness RDF data through an interactive, customised dashboard that enabled users to write a SPARQL query to select information to go in a chart. The following section concisely outlines open research challenges that Open Data Observatories' developers may encounter.

6 Research Challenges

Implementing Open Data Observatories is challenging. Interoperability, scalability, and replication are issues when integrating disparate data sources and their systems. Different designs, goals, and computing specs may be used. Integrating disparate systems can generate service conflicts, degrade data quality, lose data data provenance, and breach privacy. Data integration, context, quality, provenance, and privacy may affect Data Observatories best practises. This section addresses each challenge, showing research progress in the Open Data Observatories.

6.1 Data Integration

Data generated by the devices in the IoT require careful and timely integration. Integrating heterogeneous data can positively impact decision making. However, achieving valid integration face many challenges, as stated by many researchers such as [125, 126]. Data Observatories suffer primarily from the following challenges:

- *Interoperability*, one of the fundamentals concerns for integrating IoT data [127] across platforms [128]. It attempts to interconnect heterogeneous smart devices across heterogeneous networks. There are currently no established standards for compatibility between these devices and their applications [106]. Rather, continuing research seeks to advance the state of the art. For instance, Ullah et al. [127] built a semantic model for the healthcare sector to recommend medications with adverse effects for a variety of illnesses. Mishra et al. [129] asserted that semantic techniques might address interoperability issues by modelling data sources as ontologies and evaluating their quality, competency, integrity, and completeness. Open Data Observatories such as [75, 85, 91] used semantic modelling to address interoperability challenges when combining heterogeneous data.

- *Heterogeneity*, data maintained at Data Observatories may suffer heterogeneity issues on three levels, (i) semantic (i.e., the meaning of the data is interpreted in different ways, probably due to change in the meaning based on context and time or linguistic issues) (ii) syntactic (i.e., data with different formats, such as csv, text and image) (iii) structural (i.e., data with different storage methods) [130]. Typically, the data arrive from multiple disparate sources. Each source collects its observations from different locations at different times, making it hard for decision-makers to validate readings and deduce facts. For example, Grow Observatory [84, 83] soil-moisture sensors provide domestic and scientific(lab) data. However, the researchers [84] identified many conflicts between the two readings in terms of interpretation and structure.
- *Completeness*, Sensor data may contain missing readings on occasion owing to technical failures (e.g., power outage) or transmission issues. Complete data is of higher quality and can be used to develop more accurate machine learning models. Testing several machine learning models and using the fittest to impute missing data could be one solution.
- *Real-timeliness*, The majority of assessed observatories experience delays between data collection and dissemination. This delay is unique to each observatory and its duration may vary based on the data processing and deployment strategy. Additionally, data may be subject to ownership and confidentiality limitations. Newcastle real-time observations, for instance, are published and updated nearly every minute, whereas Bristol Urban Flows Observatory real-time project data may meet data ownership and privacy difficulties, prohibiting quick dissemination.

6.2 Data Context

Open Data Observatories, as a representation of unified IoT services [131] platforms, produce large-scale heterogeneous sensors data that contains diverse contextual information. In such dynamic environments, it is challenging to process these data with slim human intervention (e.g., make them self-configuring to adapt their behaviour at run time). Consequently, context-aware techniques can assist in understanding the situation by connecting sensors' contextual information to ambient intelligence in real-time. In other words, various IoT devices, including wearable devices, smart sensors, cameras and GPS collars, can be connected to collect context-aware information about users' surroundings. Subsequently, when analysed, the collected data from the connected devices can accomplish personalised and adaptive decision-making in context-aware applications. Although context-aware techniques can help, yet faces another set of challenges regarding data integration and privacy [132]. In computing literature, the term context-aware consumed multiple definitions and various implementations corresponding to each research question; more details were discussed in Perera et al. [133] survey.

6.3 Data Quality

Applied research defined the term data quality differently [134], a commonly used definition by Strong et al. [135] describing data quality as data that is fit for the intended purpose. Byabazaire et al. [136] and Taleb et al. [137] testified that data quality is a mature research topic in big data and databases management. However, Perez-Castillo et al. [134] claimed its youth in Smart Connected Products (SCP) [138] and the IoT. Data quality plays a significant role in IoT environments, and a sufficient quality level can build trust between the cyber and physical world [139, 136, 134]. IoT data generated in the Open Data Observatories are extensive in volume, arrive at high speed, and are heterogeneous. Moreover, the growing numbers of heterogeneous sensors and smart devices joining the IoT increased the probability of acquiring inaccurate and unfaithful data. In other words, raw data may have some missing and incorrect records due to power glitches and human errors during collection, respectively. [140, 141]. These traits pose technical challenges in securing an adequate data quality level throughout data life-cycles. Data life-cycle forms a chain of crucial steps for data to undergo, starting from the collection, curation, and processing to usage [136]. Whistle each life-cycle step may have different quality measures and evaluation, which, indeed, require careful interpretations [137], the quality and value of the same data vary from one stakeholder to another. Byabazaire et al. [136] introduced a framework to assess big data quality in the IoT using "Trust" in the absence of validation references. Neumeier et al. [142] developed a framework to automatically monitor and validate the quality of metadata in different Open Data portals. Data quality in IoT platforms must conform with standard guidelines set by professional bodies. For example, the international series of ISO/IEC 25000 [143] which deals with systems and software quality requirements and evaluation (SQuaRE) [134], and ISO 8000–60 series [144, 145] that addresses the best practices in data quality management methods. Perez-Castillo et al. [134]introduced an IoT data conceptual framework that conforms ISO 8000–62 [146] for evaluating and enhancing the quality of data Smart Connected Products (SCP) environments. In [139], Perez-Castillo et al. adapted Deming Wheel [147], "Plan-Do-Check-Act", in their method to manage data quality in sensor networks. They named the model (DAQUA-MASS) and aligned it with ISO 8000-61 [139]. Another method to evaluate data quality as reviewed by Hu et al. [148] uses data provenance, which can trace back data history and detect errors.In the context of publishing data

Table 14: Research challenges in data observatories, instances and possible solution.

Research challenges	Instances	Possible solution
Data Integration	Data are generated independently; different formats are domain-specific, lack description (metadata, dictionary and ontology).	Modelling data sources as linked data endpoints in the Resource Description Framework (RDF).
Interoperability	Diverse devices, hardware, software and communication networks.	Usage of Semantic Web Technologies (i.e., ontologies and linked data).
Data Heterogeneity	semantic, syntactic and structural differences	Usage of ontologies for formal concepts' expression and sharing.
Data Completeness	Missing data.	Usage of machine learning to impute missing data.
Data Real-timeliness	Time lag between data collection and sharing.	Implementing service agreements and privacy tools that allow instant data sharing.
Data Context	Leveraging contextual information collected from various IoT devices to understand users' surroundings and make adaptive decisions in real-time.	Modelling data context using ontologies based on the 5Ws (who, when, what, where and why) to capture the generic concepts to a higher level.
Data Quality	Fitness of devices collecting data and the generated observations.	Maintaining adequate quality assurance (i.e., devices used for data collection) and quality control (i.e., the manual and automated procedures applied to review the collected data).
Data Provenance	Identify and keep data origin.	Usage of semantic tools capable of capturing and documenting data provenance and tracing data movement.
Data Privacy	Protect sensitive and personal information.	Applying anonymisation, perturbation, cloaking and tracing data provenance.

through Open Data Observatories, the term integrity implies the importance of providing reliable and accurate data to everyone. Open Data integrity trade-off presumably unavoidable during the collection and integration processes due to many factors that may negatively impact quality assurance and control [11, 82]. Quality assurance deals with devices used for data collection, verifying their fitness, performance, and reliability. Yet, quality control concerns the manual and automated procedures applied to review the collected data [82]. Examples of these factors include sensors, devices with low quality and unsuitability for their environment, genuine human errors, low accuracy of statistical models and machine learning algorithms that process raw data, malicious security attacks, and privacy restrictions. Domain experts attempted to overcome quality concerns using several techniques. Triangulum City Dashboard [34] ensured that its data quality is fully conforming with the EU General Data Protection Regulation [149]. Grow Observatory [84, 150, 83] checked their data quality by validating the remotely sensed observations against citizens generated ground observations. Channel Coastal Observatory [81] outsourced the data quality control to ensure that the specification of each data type (e.g., Lidar, tides, waves) meets the required standards. Newcastle Urban Observatory [46] and Bristol Urban Flows Observatory [59] deployed different types of sensors at the same place and compared their records (e.g. CCTV cameras to validate a sound sensor for noise at the same building).

6.4 Data Provenance

Data provenance at Open Data Observatories compromises tracing the roots from generation and derivation of data over time. It is a conventional approach in data mining and databases systems disciplines, primarily employed in diverse cloud-based applications to assure shared data quality, integrity, and privacy [151]. Pearce [152] stated that any data arriving from any source are credible if one could locate them and identify their lineage. Open Data Observatories often supplement their data with metadata - also known as provenance data- to enable users to understand, trust and rely on the data in question. Hu et al. [148] in their recent survey, explained the difference between the two concepts *data provenance* and *provenance data*. The former is the method that records data origin and growth, while the latter refers to the information (metadata) documented by the method. Adapting data provenance methods in IoT environments is deemed challenging due to the dynamic data nature and the IoT devices computational power [151, 148]. At present, the schemes used for data provenance as discussed in [148] and [153] include blockchain-based, cryptography-based and logging -based. Hu et al. [148], for instance, extended the three-layered implementation for IoT smart services in Yang et al. [154] by injecting a middle-layer. This novel model accounts for data provenance management by integrating each layer's services. Even with the proposed model capabilities in tracking IoT data behaviour and enhancing their quality and security yet experienced some technical challenges [148]. First, data in IoT, typically, travel across multiple execution layers, process and mix recurrently by different applications, making it difficult to keep detailed track of their historical activities- besides identifying their root nodes, detecting generation and processing errors. Second, every time the data undergo a new transmission or execution, a different service will generate new metadata. This proportional increase demands larger storage space and memory, leading to difficulties updating and retrieving provenance data.

6.5 Data Privacy

The massive data that IoT devices continuously accumulate in Open Data Observatories orderly undergo collection [156], aggregation [157] and data analytics [158, 101]. In any event, part of users' sensitive data is likely to uncover,

Table 15: Pros and cons of the reviewed Open Data Observatories, Future Recommendations and Take-aways to non-urban observatories.

Data Observatory	Pros	Cons	Future Recommendation	Take-aways to non-urban Observatories
Newcastle [46]	Largest diverse datasets collection worldwide, real-time information and easy to use interface	Lack of evident research documenting the positive impact of the observatory on Newcastle city. (e.g., reduce crime rates)	Replicate projects to more cities and remote areas, locally and globally	Data ingestion, presentation and sharing
Sheffield [58]	Using of sensing vehicles and drones supported evident decision making	Lack of real-time interactive visualisation.	Replicate projects to other UK cities	Sensing vehicles and drones
Bristol [59]	Active research community with many pilot local projects	Access to real-time data faces data ownership and privacy obstacles	More research to establish data ownership and privacy tools to validate data collection, processing and public sharing.	WSN deployment and WI-FI outsourcing
Cranfield [71]	Observatory data support the teaching and learning in Cranfield University	Limited data varieties and publications	More reports to reflect the living lab initiatives	Monitoring animals using acoustic sensors
Birmingham [73]	Downloadable real-time sensors data displayed via interactive map	Limited publications	Given the busy nature of Birmingham city, more sensors coverage and analysis tool kits for the time-series data would be more beneficial for end-users	Real-time birds detector
Manchester [74, 75]	First UK source for real-time pollen-concentration data. The crowd-sourcing feature allows citizens to express their views and opinions	No social media in the platform	Replicate projects to other UK cities	Crowdsourcing and the Semantic Web approach
Triangulum [34]	Open-source applications and modern web frameworks integration	Issues with data quality and privacy	More research to establish data quality and privacy tools to validate data collection, processing and public sharing.	Multi-stakeholder partnerships, top-down support
Channel Coastal [81]	Time-series and extreme values analysis research	Outsourcing data storage may impose security concerns	Integrate more real-time monitoring systems	Extreme events analysis
Grow [84][83]	Empowers citizens and communities to have a say on soil and climate matters across Europe	Heterogeneous data integration issues and outsourcing data storage may impose security concerns	Integrate more data sources such as drones	Do-it-Yourself (DIY) sensors deployment
Linked Data Finland [85]	Applied semantic web on multiple research areas (museum, health, and environment)	No social media in the platform	More research to reflect the observatory size and its capabilities	Data documentations and validation tools
LOD IoT Health [91]	Valid use for semantic web in integrating heterogeneous IoT datasets	Privacy of health information concerns	Link the semantic sensors readings to Electronic Medical Record (EMR) systems to support clinical decisions [155]	Semantic web processing
Cyclone [36]	Multiple holistic and real-time dashboards dispersed across Australia and the Pacific Islands	Lack of direct access to data and analysis tools	Supply downloadable datasets via integrated APIs and cloud-based data analysis	Real-time data presentation

including - but not limited to personally identifiable information, financial status, health records, and lifestyle habits [156]. Short of adequate privacy protection in IoT applications could cause unwelcome privacy invasion and may harm individual welfare [101, 159, 160, 161, 162] and [163]. For example, in most cases, when users attain IoT data from Open Data platforms, they fill in a request form that takes in names, emails and addresses. If data privacy is not securely maintained, these data could fall into the wrong hands. IoT devices may also store private information– mostly with no clear notice- and transmit them to remote storage or other devices in the network [164]. Observatory owners must trust their devices and have a clear guideline on what and how much data to collect and only to hold data if necessary. Nevertheless, they must ensure that IoT data stay in its original form and is accessible with only appropriate permissions. Data privacy protection attracted researcher's attention for the past decade, with implemented solutions such as anonymisation [165, 166], perturbation [167], cloaking [156, 165] and data provenance [148], yet, implementing privacy-preserving systems in IoT with optimally safe data usage remains a challenge [101]. For instance, Reda et al. [91] claimed that one of their portal limitations was the lack of data privacy best practices, which in turn, may put sensitive health information at risk of disclosure. Liu et al. [156] introduced a novel IoT data collection method that protects individual information privacy. The mechanism works by cloaking the data source from consumers. Perera et al. [168] implemented a tool that can raise awareness of privacy shortcomings in IoT applications. Newcastle Urban Observatory [46] has a privacy policy in place to deal with data collection and dissemination [169]. The policy complied with the requirements of the EU General Data Protection Regulation (GDPR) [79] and shared at Newcastle urban observatory website [46]. Whistle visiting the urban observatories website, the user IP address is identified and

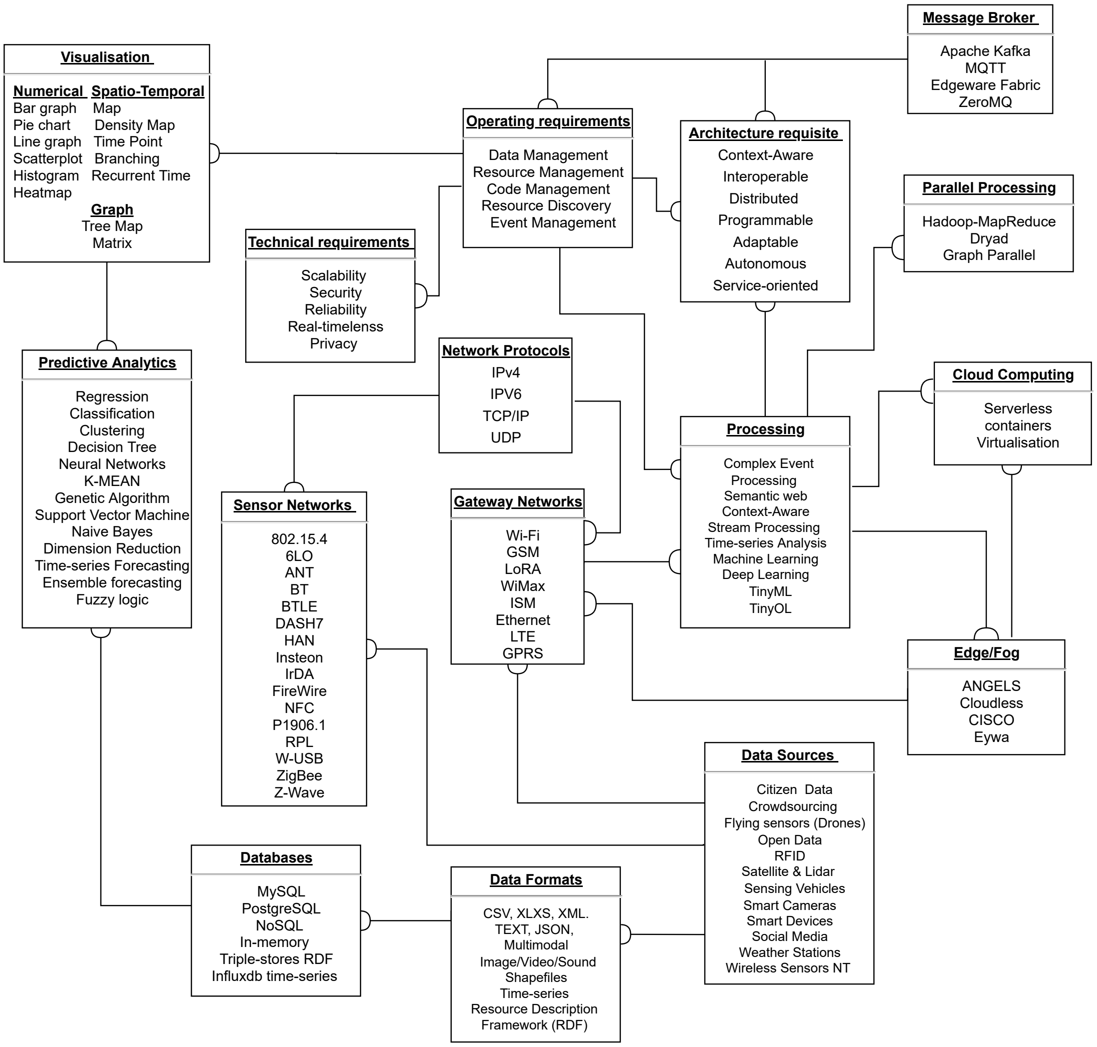


Figure 8: Taxonomy for Data Observatories Features

other information such as the browsing referral. These data are stored for a month for administration purposes. To acquire data from the urban observatories, users have to create email accounts. This information will not be shared with third parties and marketing companies without the users acceptance. Here, user information showing data requests is kept for one year before anonymising for funders reports and statistics.

- Privacy of sensors data collection at Newcastle Helix, a university building, accommodates nearly 3,000 various sensors for performance evaluation. These sensors can generate personal data when observing areas with personal spaces (e.g., private offices). Personal data here will be restricted from public sharing through APIs. However, under certain circumstances, given that ethical approval is obtained, personal data - excluding names are used for scientific and analysis purposes.
- Privacy of sensors data collection in Newcastle upon Tyne, Here, hundreds of sensors monitor metrics such as air quality, electricity and gas usage. Personal data are kept private and anonymised. There are also data extracted from images using machine learning. Only the non-personal data are aggregated. Multimedia data used during training and testing the machine learning models undergo strict supervision, then disposed of at the end of the project.

Another GDPR-compliant privacy policy[113] released by the Grow Observatory[83, 84] details the storage, processing, and sharing of customer data. Compared to [46], the Grow Observatory policy contains more information. For example, rigorous classification for the users' information and scientific data as well as definitions of sensitive data, personally

Table 16: Star rating based on the 5-star models for Data forms [19], Data Engagement [20], and Open Data Portals [21].

Data Observatory	Data Forms	Data Engagement	Data Portal
Newcastle [46]	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆
Sheffield [58]	⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆	⋆
Bristol [59]	⋆ ⋆ ⋆	⋆ ⋆ ⋆	⋆
Cranfield [71]	⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆
Birmingham [73]	⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆
Manchester [74, 75]	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆
Triangulum [34]	⋆ ⋆ ⋆	⋆ ⋆ ⋆	⋆
Channel Coastal [81]	⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆
Linked Data Finland [87] [85]	⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆
Grow Observatory [84][83]	⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆
Cyclone [36]	⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆
LOD IoT Health [91]	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆	⋆

identifiable information, and aggregated data. In addition, the policy specifies tight regulations for website usage and defines the contents that are permitted and banned for uploading.

7 Conclusion

With the rapid expansion of Internet of Things (IoT) devices and data processing tools, mostly in smart cities, a number of unanswered questions arise regarding their current condition, use cases, and future development [105]. In the state of development, it is vital to have ongoing support in terms of funds and scientific research. Consequently, suitable data management solutions can be implemented to provide smooth integration and an optimal user experience. Open Data Observatories provide answers by coordinating the management and dissemination of data. To primarily serve the public interest through engaging citizens, promoting openness, and facilitating informed decision-making. This survey investigated Open Data as the primary data source in Open Data Observatories, and then examined twelve urban observatories to see what can be replicated in non-urban locations. Our findings revealed intriguing information regarding the Open Data Observatories under consideration. The Newcastle Urban Observatory integrated the most comprehensive data sets, accessible sources, and real-time updates from social media into a single, user-friendly interface. As such, Sheffield Urban Observatory's sensing vehicles and drones stood out. Grow Observatory delivered on citizens' engagement. Simultaneously, Channel Coastal Observatory and Cyclones Dashboards emphasised the forecasting of natural disasters. On the one hand, the Manchester Urban Observatory implemented semantic web and crowdsourcing features, resulting in a higher star rating. Linked Data Finland, on the other hand, went above and above by embracing quick documentation and validation technologies. We propose that the characteristics of the examined observatories may inspire the establishment of fresh observatories, whether in urban or rural settings. We assessed the reviewed observatories in light of our findings by highlighting their benefits and drawbacks. In the same table 15, we summarised a few functional elements to be replicated in non-urban observatories and proposed future recommendations. As a result, we drew up a taxonomy 8 to categorise the key elements that guided the development of data observatories. Last but not least, we awarded stars to each Observatory, as shown in table 16, based solely on the aforementioned 5-star models. The 5-star models compared the data forms, engagement and data portals. For the data forms, the fifth star was awarded to the comprehensive Linked Open Data (LOD), whilst for engagement, external collaboration and citizens participation were deemed to be the most effective. The fifth star went to the portal for data portals that achieved interoperability and provided data provenance, governance, and quality assurance. In a nutshell, replicating the features of urban data observatories in non-urban environments necessitates top-down support and bottom-up data systems.

References

[1] Hawkr | Urban Observatory Projects, 2020.

[2] Gareth W Young, Rob Kitchin, and Jeneen Naji. Building City Dashboards for Different Types of Users. *The Journal of urban technology*, pages 1–21, 2020.

[3] K K Lwin, Y Sekimoto, W Takeuchi, and K Zettsu. City Geospatial Dashboard: IoT and Big Data Analytics for Geospatial Solutions Provider in Disaster Management. In *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–4, 12 2019.

[4] Rodrigo Tapia-McClung. Exploring the use of a spatio-temporal city dashboard to study criminal incidence: A case study for the Mexican state of aguascalientes. *Sustainability (Switzerland)*, 12(6), 2020.

[5] Samuel Stehle and Rob Kitchin. Real-time and archival data visualisation techniques in city dashboards. *International Journal of Geographical Information Science*, 34(2):344–366, 2020.

[6] Christopher Pettit, Scott N Lieske, and Murad Jamal. CityDash: Visualising a changing city using open data. In *International Conference on Computers in Urban Planning and Urban Management*, pages 337–353. Springer, 2017.

[7] Luigi Atzori, Antonio Iera, and Giacomo Morabito. Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Networks*, 56:122–140, 2017.

[8] Atheer Aljeraisy, Omer Rana, and Charith Perera. A Systematic Analysis of Privacy Laws and Privacy by Design Schemes for the Internet of Things: A Developer's Perspective. 2020.

[9] Alex Mavromatis, Sam Gunner, Theo Tryfonas, and Dimitra Simeonidou. Dynamic cloud service management for scalable internet of things applications. In *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1978–1982, 2019.

[10] Lihong Jiang, Li Da Xu, Hongming Cai, Zuhai Jiang, Fenglin Bu, and Boyi Xu. An IoT-Oriented data storage framework in cloud computing platform. *IEEE Transactions on Industrial Informatics*, 10(2):1443–1451, 2014.

[11] Meiyi Ma, Sarah M. Preum, Mohsin Y. Ahmed, William Tärneberg, Abdeltawab Hendawi, and John A. Stankovic. Data sets, modeling, and decision making in smart cities: A survey. *ACM Transactions on Cyber-Physical Systems*, 4(2), 2019.

[12] Vishanth Weerakkody, Zahir Irani, Kawal Kapoor, Uthayasankar Sivarajah, and Yogesh K. Dwivedi. Open data and its usability: an empirical view from the Citizen's perspective. *Information Systems Frontiers*, 19(2):285–300, 2017.

[13] Dan Hughes. What is open data and how do you use it? *The Estates Gazette*, page 39, 2018.

[14] Alberto Abella, Marta Ortiz-de Urbina-Criado, and Carmen De-Pablos-Heredero. The process of open data publication and reuse. *Journal of the Association for Information Science and Technology*, 70(3):296–300, 2019.

[15] Philipp Lämmel, Benjamin Dittwald, Lina Bruns, Nikolay Tcholtchev, Yuri Glikman, Silke Cuno, Mathias Flügge, and Ina Schieferdecker. Metadata harvesting and quality assurance within open urban platforms. *J. Data and Information Quality*, 12(4), oct 2020.

[16] Jeremy V Kepner. *Mathematics of big data : spreadsheets, databases, matrices, and graphs*. MIT Lincoln laboratory series. 2018.

[17] Vijayakumar Nanjappan, Hai Ning Liang, Wei Wang, and Ka L. Man. *Big Data: A Classification of Acquisition and Generation Methods*. Elsevier Inc., 2017.

[18] Gabriela Viale Pereira, Marie Anne Macadar, Edimara M Luciano, and Maurício Gregianin Testa. Delivering public value through open government data initiatives in a smart city context. *Information systems frontiers*, 19(2):213–229, 2016.

[19] Berners-Lee Tim. 5-star Open Data, 2006.

[20] Tim Davies. Open Data Engagement: Exploring the engagement dimensions of open data, 2012.

[21] P Colpaert and S Joye. The 5 stars of open data portals. In *Proceedings of the 7th International Conference on Methodologies, Technologies and Tools Enabling E-Government*, pages 61–67, 2013.

[22] Eunice J Yuan, Chia-An Hsu, Wui-Chiang Lee, Tzeng-Ji Chen, Li-Fang Chou, and Shinn-Jang Hwang. Where to buy face masks? Survey of applications using Taiwan's open data in the time of coronavirus disease 2019. *Journal of the Chinese Medical Association : JCMA*, 83(6):557–560, 6 2020.

[23] Scott Hawken, Hoon Han, and Christopher Pettit. Introduction: Open Data and the Generation of Urban Value. In Scott Hawken, Hoon Han, and Chris Pettit, editors, *Open Cities | Open Data: Collaborative Cities in the Information Era*, pages 1–25. Springer Singapore, Singapore, 2020.

[24] The 8 Principles of Open Government Data (OpenGovData.org), 2014.

[25] Sunlight Foundation. Ten Principles for Opening Up Government Information. *Sunlight Foundation*, (October 2007):3, 2010.

[26] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, W. Hall, and M. C. Schraefel. Open Government Data and the Linked Data Web: Lessons from data. gov. uk. *IEEE Intelligent Systems*, 27(3):16–24, 2012.

[27] Great Britain. H.M. Treasury. *Putting the frontline first : smarter government*. Stationery Office, 2009.

[28] Victoria Wang and David Shepherd. Exploring the extent of openness of open government data - A critique of open government datasets in the UK. *Government Information Quarterly*, 37(1):101405, 2020.

[29] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Soren Auer. A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418, 2015.

[30] Noor Huijboom and Tijs Van den Broek. Open data: an international comparison of strategies. *European journal of ePractice*, 12(1):4–16, 2011.

[31] Maria-Lluisa Marsal-Llacuna, Joan Colomer-Llinàs, and Joaquim Meléndez-Frigola. Lessons in urban monitoring taken from sustainable and livable cities to better address the Smart Cities initiative. *Technological Forecasting and Social Change*, 90:611–622, 2015.

[32] Harvey Miller and Kristin Tolle. Big Data for Healthy Cities: Using Location-Aware Technologies, Open Data and 3D Urban Models to Design Healthier Built Environments. *Built Environment*, 42:441–456, 2016.

[33] N. S. Widodo, S. A. Akbar, and A. Rahman. Robot Operating System (ROS) Compatible Low Cost Rotating Light Detection and Ranging (Lidar) Design. *IOP Conference Series: Materials Science and Engineering*, 384(1), 2018.

[34] Mina Farmanbar and Chunming Rong. Triangulum City Dashboard: An Interactive Data Analytic Platform for Visualizing Smart City Performance. *Processes*, 8(2):250, 2 2020.

[35] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020.

[36] Ian Tilley and Christopher Pettit. A Dashboard for the Unexpected: Open Data for Real-Time Disaster Response. In Scott Hawken, Hoon Han, and Chris Pettit, editors, *Open Cities | Open Data: Collaborative Cities in the Information Era*, pages 265–286. Springer Singapore, Singapore, 2020.

[37] About Us - Urban Flows Observatory, 2018.

[38] Alexis Madrigal and Robinson Meyer. The COVID Tracking Project, 2020.

[39] Urban Observatory monitors Newcastle with new generation of environmental sensors, 2020.

[40] Unai Lopez, Jeffrey Morgan, Kathryn Jones, Omer Rana, Timothy Edwards, and Fabio Grigoletto. Enabling citizen science in rural environments with IoT and mobile technologies, 2019.

[41] M. Woods, D. Hemment, R. Ajates, A. Cobley, A. Xaver, and G. Konstantakopoulos. GROW Citizens' Observatory: Leveraging the power of citizens, open data and technology to generate engagement, and action on soil policy and soil moisture monitoring. *IOP Conference Series: Earth and Environmental Science*, 509(1):10–12, 2020.

[42] Yiheng Chen and Dawei Han. Water quality monitoring in smart city: A pilot project. *Automation in Construction*, 89:307–316, 2018.

[43] Parvaneh Asghari, Amir Masoud Rahmani, and Hamid Haj Seyyed Javadi. Internet of Things applications: A systematic review, 2019.

[44] Gregory Dobler, Federica B Bianco, Mohit S Sharma, Andreas Karpf, Julien Baur, Masoud Ghandehari, Jonathan S. Wurtele, and Steven E Koonin. The Urban Observatory: a Multi-Modal Imaging Platform for the Study of Dynamics in Complex Urban Systems. 2019.

[45] Luke Smith and Mark Turner. Building the Urban Observatory : Engineering the largest set of publicly available real-time environmental urban data in the UK. 21:10456, 2019.

[46] Urban Observatory, 2015.

[47] INSIGHTS Virtual Lectures: Quantifying the impact of Covid-19 on city systems - YouTube, 2020.

[48] Phil James, Richard Dawson, Stuart Barr, and Neil Harris. Centre for Earth Systems Engineering Research Urban Observatory Mapping our Future Cities http://uoweb1.ncl.ac.uk.

[49] Chih Chieh Hung and Chu Cheng Hsieh. *Big Data Management on Wireless Sensor Networks*. Elsevier Inc., 2017.

[50] Apache Kafka, 2021.

[51] DfT Transport Technology Innovation Grant | Urban Observatory Projects, 2020.

[52] Urban Observatory - Healthy Schools, 2019.

[53] Office Wellbeing | Urban Observatory Projects, 2020.

[54] Virtual reality BIM and BMS | Urban Observatory Projects, 2020.

[55] REBUSCOV | Urban Observatory Projects, 2020.

[56] Flood Prepared | Urban Observatory Projects, 2020.

[57] Philip James, Ronnie Das, Agata Jalosinska, and Luke Smith. Smart cities and a data-driven response to COVID-19, 2020.

[58] Home - Urban Flows Observatory, 2020.

[59] Bristol Infrastructure Collaboratory | Faculty of Engineering | University of Bristol, 2019.

[60] Home - DEPRECATED Old FIWARE DataModels, 2021.

[61] Raffaele Zinno, Serena Artese, Gabriele Clausi, Floriana Magar, Sebastiano Meduri, Angela Miceli, and Assunta Venneri. Structural health monitoring (shm). In *The Internet of Things for Smart Urban Ecosystems*, pages 225–249. Springer, 2019.

[62] Sam Gunner, Paul J Vardanega, Theo Tryfonas, John H G Macdonald, and R Eddie Wilson. Rapid deployment of a wsn on the clifton suspension bridge, uk. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, 170(3):59–71, 2017.

[63] Md Anam Mahmud, Kyle Bates, Trent Wood, Ahmed Abdelgawad, and Kumar Yelamarthi. A complete internet of things (iot) platform for structural health monitoring (shm). In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, pages 275–279, 2018.

[64] Sukun Kim, Shamim Pakzad, David Culler, James Demmel, Gregory Fenves, Steve Glaser, and Martin Turon. Wireless sensor networks for structural health monitoring. In *Proceedings of the 4th international conference on Embedded networked sensor systems*, pages 427–428, 2006.

[65] MSCL, 2021.

[66] GitHub - influxdata/influxdb: Scalable datastore for metrics, events, and real-time analytics, 2021.

[67] Grafana: The open observability platform | Grafana Labs, 2021.

[68] Travel Exhibition | Faculty of Engineering | University of Bristol, 2022.

[69] David Nepomuceno, Theo Tryfonas, and PJ Vardanega. Residential damp detection with temperature and humidity urban sensing. In *International Conference on Smart Infrastructure and Construction 2019 (ICSIC) Driving data-informed decision-making*, pages 605–611. ICE Publishing, 2019.

[70] Cranfield Urban Observatory , 2021.

[71] Cranfield University's Living Laboratory and Urban Observatory - About, 2021.

[72] Jacqueline Hannam and Carolyn Nandozi. Using low cost sensors to assess soil temperature response to summer heatwaves in urban greenspaces . (May):18007, 2020.

[73] Home | Birmingham Urban Observatory, 2018.

[74] Manchester Urban Observatory, 2015.

[75] Manchester-I, 2017.

[76] Laura Keast, Lindsay Bramwell, Kamal Jyoti Maji, Judith Rankin, and Anil Namdeo. Air quality outside schools in newcastle upon tyne, uk: An investigation into no2 and pm concentrations and pm respiratory deposition. *Atmosphere*, 13(2), 2022.

[77] Markus Lanthaler and Christian Gütl. On using JSON-LD to create evolvable RESTful services. WS-REST '12, pages 25–32. ACM, 2012.

[78] Bham Urban Obs Data, 2022.

[79] General Data Protection Regulation (GDPR) – Official Legal Text, 2021.

[80] Stilianos Contarinis, Athanasios Pallikaris, and Byron Nakos. The Value of Marine Spatial Open Data Infrastructures-Potentials of IHO S-100 Standard t Become the Universal Marine Data Model. *Journal of Marine Science and Engineering*, 8(8):564, 2020.

[81] Channel Coastal Observatory - Porthleven, 2019.

[82] Travis Mason and Thomas Dhoop. Cover photograph: Datawell Directional Waverider Mk III in Weymouth Bay Photo courtesy of Fugro GB Marine Limited National Network of Regional Coastal Monitoring Programmes of England Quality Assurance & Quality Control of Wave Data. 2017.

[83] About the Project – The GROW Observatory, 2021.

[84] Karoly Zoltan Kovács, Drew Hemment, Mel Woods, Naomi K. van der VELDEN, Angelika Xaver, Rianne H. Gi Esen, Victoria J. Burton, Natalie L. Garrett, Luca Zappa, Deborah Long, Endre Dobos, and Rastislav Skalsky. Citizen observatory based soil moisture monitoring - The GROW example. *Hungarian Geographical Bulletin*, 68(2):119–139, 2019.

[85] Eero Hyvönen, Jouni Tuominen, Miika Alonen, and Eetu Mäkelä. Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In Valentina Presutti, Eva Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events*, pages 226–230, Cham, 2014. Springer International Publishing.

[86] V. Meghana, K. Ashika, B. Namitha, and Nalini Sampath. Design of an Integrated System for Monitoring Weather and Traffic Based on Internet of Things. *Lecture Notes on Data Engineering and Communications Technologies*, 26(2):942–951, 2019.

[87] Linked (Open) Data Finland - Living Laboratory Data Service for the Semantic Web, 2021.

[88] Tarcisio Mendes de Farias and Christophe Dessimoz. Enhancing interoperable datasets with virtual links. *arXiv preprint arXiv:1906.01950*, 2019.

[89] Tarcisio Mendes de Farias, Kurt Stockinger, and Christophe Dessimoz. Voidext: Vocabulary and patterns for enhancing interoperable datasets with virtual links. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 607–625. Springer, 2019.

[90] David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. RDF 1.1 Turtle, 2014.

[91] Roberto Reda and Antonella Carbonaro. Design and development of a linked open data-based web portal for sharing IoT health and fitness datasets. *ACM International Conference Proceeding Series*, pages 43–48, 2018.

[92] Antonella Carbonaro, Filippo Piccinini, and Roberto Reda. Integrating heterogeneous data of healthcare devices to enable domain data management. *Journal of E-Learning and Knowledge Society*, 14(1):45–56, 2018.

[93] Roberto Reda, Filippo Piccinini, and Antonella Carbonaro. Towards consistent data representation in the IoT healthcare landscape. *ACM International Conference Proceeding Series*, 2018-April:5–10, 2018.

[94] Kanishk Chaturvedi and Thomas H Kolbe. InterSensor Service: Establishing Interoperability over Heterogeneous Sensor Observations and Platforms for Smart Cities. pages 1–8. IEEE, 2018.

[95] Charles Carter and Chris Rushton. Road Transport and Air Quality. *The Internet of Things*, pages 189–206, 2020.

[96] Mogens Jin Pedersen and Nathan Favero. Social Distancing during the COVID-19 Pandemic: Who Are the Present and Future Noncompliers? *Public Administration Review*, 80(5):805–814, 2020.

[97] Amir Javed, Pete Burnap, Matthew L. Williams, and Omer F. Rana. Emotions behind drive-by download propagation on twitter. *ACM Trans. Web*, 14(4), aug 2020.

[98] COSMOS – Social Data Science Lab, 2022.

[99] Mohit Taneja, Nikita Jalodia, John Byabazaire, Alan Davy, and Cristian Olariu. SmartHerd management: A microservices-based fog computing-assisted IoT platform towards data-driven smart dairy farming. *Software - Practice and Experience*, 49(7):1055–1078, 2019.

[100] Yun-Yi Zhang, Kai Kang, Jia-Rui Lin, Jian-Ping Zhang, and Yi Zhang. Building information modeling-based cyber-physical platform for building performance monitoring. *International Journal of Distributed Sensor Networks*, 16(2), 2020.

[101] Eugene Siow, Thanassis Tiropanis, and Wendy Hall. Analytics for the internet of things: A survey. *ACM Computing Surveys*, 51(4), 2018.

[102] Mahdi Fahmideh and Didar Zowghi. An exploration of IoT platform development, 2020.

[103] Yassine Chahid, Mohamed Benabdellah, and Abdelmalek Azizi. Internet of things protocols comparison, architecture, vulnerabilities and security: State of the art. *ACM International Conference Proceeding Series*, pages 0–5, 2017.

[104] Srikar Meka and Benedito Fonseca. Improving route selections in zigbee wireless sensor networks. *Sensors (Switzerland)*, 20(1):1–34, 2020.

[105] *Management of IOT Open Data Projects in Smart Cities*. 2021.

[106] Ammar Gharaibeh, Mohammad A. Salahuddin, Sayed Jahed Hussini, Abdallah Khreishah, Issa Khalil, Mohsen Guizani, and Ala Al-Fuqaha. Smart cities: A survey on data management, security, and enabling technologies. *IEEE Communications Surveys Tutorials*, 19(4):2456–2501, 2017.

[107] Margaret Rouse. What is data management and why is it important?, 2020.

[108] Lena Wiese. *Advanced data management: For SQL, NoSQL, cloud and distributed databases*. 2015.

[109] Amir Sinaeepourfard, Jordi Garcia, Xavier Masip-Bruin, and Eva Marin-Tordera. Data Preservation through Fog-to-Cloud (F2C) Data Management in Smart Cities. In *2018 IEEE 2nd International Conference on Fog and Edge Computing (ICFEC)*, pages 1–9. IEEE, 2018.

[110] Khushboo Kalia and Neeraj Gupta. Analysis of hadoop MapReduce scheduling in heterogeneous environment. *Ain Shams Engineering Journal*, 2020.

[111] Tom (Tom E.) White. *Hadoop : the definitive guide*. First edit edition, 2009.

[112] M. Mazhar Rathore, Anand Paul, Won Hwa Hong, Hyun Cheol Seo, Imtiaz Awan, and Sharjil Saeed. Exploiting IoT and big data analytics: Defining Smart Digital City using real-time urban data, 2018.

[113] Policies and Terms | GROW Observatory, 2021.

[114] Welcome | Geo-Wiki, 2021.

[115] Saravanan Thirumuruganathan, Mayuresh Kunjir, Mourad Ouzzani, and Sanjay Chawla. Automated annotations for ai data and model transparency. *ACM Journal of Data and Information Quality (JDIQ)*, 14(1):1–9, 2021.

[116] Adnan Akbar, Abdullah Khan, Francois Carrez, and Klaus Moessner. Predictive analytics for complex iot data streams. *IEEE Internet of Things Journal*, 4(5):1571–1582, 2017.

[117] Sanjay Gaur, Darshanaben D Pandya, and Deepika Soni. Closest Fit Approach Through Linear Interpolation to Recover Missing Values in Data Mining. In *Fourth International Congress on Information and Communication Technology*, volume 1041 of *Advances in Intelligent Systems and Computing*, pages 513–521. Springer Singapore, Singapore, 2019.

[118] Hsiao-Fang Yang, Chia-Hou Kay Chen, and Kuei-Ling Belinda Chen. Using Big Data Analytics and Visualization to Create IoT-enabled Science Park Smart Governance Platform. In Fiona Fui-Hoon Nah and Keng Siau, editors, *HCI in Business, Government and Organizations. Information Systems and Analytics*, pages 459–472, Cham, 2019. Springer International Publishing.

[119] Michael Evans, Dragomir Yankov, Pavel Berkhin, Pavel Yudin, Florin Teodorescu, and Wei Wu. LiveMaps: Converting Map Images into Interactive Maps. SIGIR '17, pages 897–900. ACM, 2017.

[120] Anurag Srivastava. *Mastering Kibana 6. x: Visualize Your Elastic Stack Data with Histograms, Maps, Charts, and Graphs*. Packt Publishing, Limited, Birmingham, 2018.

[121] Bruce B Frey. The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, 2018.

[122] Lorenzo Monti, Catia Prandi, and Silvia Mirri. IoT and data visualization to enhance hyperlocal data in a smart campus context. *ACM International Conference Proceeding Series*, pages 1–6, 2018.

[123] Ana Lavalle, Miguel A. Teruel, Alejandro Maté, and Juan Trujillo. Fostering sustainability through visualization techniques for real-time IoT data: A case study based on gas turbines for electricity production. *Sensors (Switzerland)*, 20(16):1–19, 2020.

[124] Ana Lavalle, Miguel A. Teruel, Alejandro Maté, and Juan Trujillo. Improving sustainability of smart cities through visualization techniques for Big Data from iot devices. *Sustainability (Switzerland)*, 12(14), 2020.

[125] Ana Maria de Carvalho Moura, Fabio Porto, Vania Vidal, Regis Pires Magalhães, Macedo Maia, Maira Poltosi, and Daniele Palazzi. A semantic integration approach to publish and retrieve ecological data. *International Journal of Web Information Systems*, 11(1):87–119, jan 2015.

[126] Maggi Bansal, Inderveer Chana, and Siobhán Clarke. A survey on iot big data: Current status, 13 v's challenges, and future directions. *ACM Comput. Surv.*, 53(6), dec 2020.

[127] Farhan Ullah, Muhammad Asif Habib, Muhammad Farhan, Shehzad Khalid, Mehr Yahya Durrani, and Sohail Jabbar. Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare. *Sustainable Cities and Society*, 34(June):90–96, 2017.

[128] Mahda Noura, Mohammed Atiquzzaman, and Martin Gaedke. Interoperability in internet of things: Taxonomies and open challenges. *Mobile Networks and Applications*, 24:796–809, 2019.

[129] Sanju Mishra and Sarika Jain. Ontologies as a semantic model in IoT. *International Journal of Computers and Applications*, 42(3):233–243, 2020.

[130] Harshana Liyanage, Paul Krause, and Simon de Lusignan. Using ontologies to improve semantic interoperability in health data. *BMJ Health & Care Informatics*, 22(2):309–315, 2015.

[131] K H K Reddy, R K Behera, A Chakrabarty, and D S Roy. A Service Delay Minimization Scheme for QoS-Constrained, Context-Aware Unified IoT Applications. *IEEE Internet of Things Journal*, 7(10):10527–10534, 10 2020.

[132] Hector John T Manaligod, Michael Joseph S Diño, Supratip Ghose, and Jungsoo Han. Context computing for internet of things. *Journal of Ambient Intelligence and Humanized Computing*, 11(4):1361–1363, 2020.

[133] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. Context aware computing for the internet of things: A survey. *IEEE Communications Surveys and Tutorials*, 16(1):414–454, 2014.

[134] R Perez-Castillo, A G Carretero, M Rodriguez, I Caballero, M Piattini, A Mate, S Kim, and D Lee. Data Quality Best Practices in IoT Environments. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, pages 272–275, 2018.

[135] Diane M Strong, Yang W Lee, and Richard Y Wang. Data Quality in Context. *Commun. ACM*, 40(5):103–110, 5 1997.

[136] J Byabazaire, G O'Hare, and D Delaney. Data Quality and Trust : A Perception from Shared Data in IoT. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, 6 2020.

[137] I Taleb, M A Serhani, and R Dssouli. Big Data Quality: A Survey. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 166–173, 7 2018.

[138] Pai Zheng, Xun Xu, and Chun Hsien Chen. A data-driven cyber-physical approach for personalised smart, connected product co-development in a cloud-based environment. *Journal of Intelligent Manufacturing*, 31(1):3–18, 2020.

[139] Ricardo Perez-Castillo, Ana Carretero, Ismael Caballero, Moises Rodriguez, Mario Piattini, Alejandro Mate, Sunho Kim, and Dongwoo Lee. DAQUA-MASS: An ISO 8000-61 Based Data Quality Management Methodology for Sensor Data. *Sensors (Basel, Switzerland)*, 18(9):3105, 2018.

[140] Jai Prakash Bhati, Dimpal Tomar, and Satvik Vats. Examining Big Data Management Techniques for Cloud-Based IoT Systems. pages 164–191, 2017.

[141] Justin M Johnson and Taghi M Khoshgoftaar. A survey on classifying big data with label noise. *J. Data and Information Quality*, oct 2021. Just Accepted.

[142] Sebastian Neumaier, Jurgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality*, 8(1), 2016.

[143] ISO/IEC. ISO/IEC 25000:2014 - Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE. Technical report, 2014.

[144] Mustafa Aljumaili. *Data quality assessment: Applied in maintenance*. PhD thesis, Luleå tekniska universitet, 2016.

[145] Mustafa Aljumaili, Ramin Karim, and Phillip Tretten. Quality of streaming data in condition monitoring using iso 8000. In *Current Trends in Reliability, Availability, Maintainability and Safety*, pages 703–715. Springer, 2016.

[146] ISO. Data quality - Part 62: Data quality management: Organizational process maturity assessment: Application of standards relating to process assessment. Technical report, 2018.

[147] Corinne N Johnson. The benefits fo PDCA. *Quality Progress*, 35(5):120, 2002.

[148] Rui Hu, Zheng Yan, Wenxiu Ding, and Laurence T. Yang. A survey on data provenance in IoT. *World Wide Web*, 23(2):1441–1463, 2020.

[149] *EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide - Second edition*. IT Governance Publishing, Ely, 2nd ed. edition, 2017.

[150] Unai Alegre, Juan Carlos Augusto, and Tony Clark. Engineering context-aware systems and applications: A survey. *Journal of Systems and Software*, 117:55–83, 2016.

[151] Adel Alkhalil and Rabie A. Ramadan. IoT Data Provenance Implementation Challenges. *Procedia Computer Science*, 109(2014):1134–1139, 2017.

[152] Henry Pearce. The (UK) Freedom of Information Act's disclosure process is broken: where do we go from here? *Information and Communications Technology Law*, 29(3):354–390, 2020.

[153] Yuan Zhang, Chunxiang Xu, and Xuemin Sherman Shen. Secure Data Provenance. In *Data Security in Cloud Storage*, pages 119–141. Springer Singapore, Singapore, 2020.

[154] Zhihong Yang, Yufeng Peng, Yingzhao Yue, Xiaobo Wang, Yu Yang, and Wenji Liu. Study and application on the architecture and key technologies for IOT. pages 747–751. IEEE, 2011.

[155] Travis R Goodwin and Sanda M Harabagiu. Knowledge representations and inference techniques for medical question answering. *ACM transactions on intelligent systems and technology (TIST)*, 9(2):1–26, 2017.

[156] Yi Ning Liu, Yan Ping Wang, Xiao Fen Wang, Zhe Xia, and Jing Fang Xu. Privacy-preserving raw data collection without a trusted authority for IoT. *Computer Networks*, 148:340–348, 2019.

[157] Tong Li, Chongzhi Gao, Liaoliang Jiang, Witold Pedrycz, and Jian Shen. Publicly verifiable privacy-preserving aggregation and its application in IoT. *Journal of Network and Computer Applications*, 126(October 2018):39–44, 2019.

[158] Mohsen Marjani, Fariza Nasaruddin, Abdullah Gani, Ahmad Karim, Ibrahim Abaker Targio Hashem, Aisha Siddiqa, and Ibrar Yaqoob. Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. *IEEE access*, 5:5247–5261, 2017.

[159] Charith Perera, Yongrui Qin, Julio C Estrella, Stephan Reiff-Marganiec, and Athanasios V Vasilakos. Fog computing for sustainable smart cities: A survey. *ACM Computing Surveys*, 50(3), 2017.

[160] Hossein Ahmadi, Goli Arji, Leila Shahmoradi, Reza Safdari, Mehrbakhsh Nilashi, and Mojtaba Alizadeh. *The application of internet of things in healthcare: a systematic literature review and classification*, volume 18. Springer Berlin Heidelberg, 2019.

[161] Chao Li and Balaji Palanisamy. Privacy in Internet of Things: From Principles to Technologies. *IEEE Internet of Things Journal*, 6(1):488–505, 2019.

[162] P Ravi Kumar, Au Thien Wan, and Wida Susanty Haji Suhaili. Exploring Data Security and Privacy Issues in Internet of Things Based on Five-Layer Architecture. *International journal of communication networks and information security*, 12(1):108–121, 2020.

[163] George Dunea. Privacy concerns. *BMJ*, 329(7464):519, 2004.

[164] Lena Wiese, Tim Waage, and Michael Brenner. CloudDBGuard: A framework for encrypted data storage in NoSQL wide column stores. *Data and Knowledge Engineering*, 126(August 2019):101732, 2020.

[165] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. -diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[166] Ali Khoshgozaran, Cyrus Shahabi, and Houtan Shirani-Mehr. Location privacy: Going beyond K-anonymity, cloaking and anonymizers. *Knowledge and Information Systems*, 26(3):435–465, 2011.

[167] Thanga S. Revathi, N. Ramaraj, and S. Chithra. Tracy-Singh Product and Genetic Whale Optimization Algorithm for Retrievable Data Perturbation for Privacy Preserved Data Publishing in Cloud Computing. *Computer Journal*, 63(2):239–253, 2020.

[168] Charith Perera, Mahmoud Barhamgi, and Massimo Vecchio. Envisioning Tool Support for Designing Privacy-Aware Internet of Thing Applications. 4500, 2017.

[169] Privacy policy | Urban Observatory, 2021.