Accepted manuscript

As a service to our authors and readers, we are putting peer-reviewed accepted manuscripts (AM) online, in the Ahead of Print section of each journal web page, shortly after acceptance.

Disclaimer

The AM is yet to be copyedited and formatted in journal house style but can still be read and referenced by quoting its unique reference number, the digital object identifier (DOI). Once the AM has been typeset, an 'uncorrected proof' PDF will replace the 'accepted manuscript' PDF. These formatted articles may still be corrected by the authors. During the Production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal relate to these versions also.

Version of record

The final edited article will be published in PDF and HTML and will contain all author corrections and is considered the version of record. Authors wishing to reference an article published Ahead of Print should quote its DOI. When an issue becomes available, queuing Ahead of Print articles will move to that issue's Table of Contents. When the article is published in a journal issue, the full reference should be cited in addition to the DOI.

Submitted: 12 January 2023

Published online in 'accepted manuscript' format: 20 September 2023

Manuscript title: Data Integration for Digital Twins in the built environment based on federated data models

Authors: Jorge Merino, Xiang Xie, Nicola Moretti, Janet Yoon Chang, Ajith Parlikad

Affiliation: Centre for Digital Built Britain, University of Cambridge, Cambridge, United Kingdom..

Corresponding author: Jorge Merino, 17 Charles Babbage Road, Alan Reece Building, Institute for Manufacturing, CB3 0FS, Cambridge, United Kingdom.

E-mail: jm2210@cam.ac.uk

Accepted manuscript
doi: 10.1680/jsmic.23.00002

Abstract

Improving efficiency of operations is a major challenge in Facility Management given the limitations of outsourcing individual building functions to third-party companies. The status of each building function is isolated in siloes which are controlled by these third-party companies. Companies provide access to aggregated information in the form of reports through web portals, emails, or bureaucratic processes. Digital Twins represent an emerging approach to return awareness and control to facility managers by automating all levels of information access (from granular data to defined KPIs and reports) and actuation. This paper proposes a low-latency data integration method that supports actuation and decision making in Facility Management, including construction, operations and maintenance data, and Internet of Things. The method uses federated data models and semantic web ontologies, and it is implemented within a data lake architecture with connections to siloed data to keep the delegation of responsibilities of data owners. A case study in the Alan Reece building (Cambridge, United Kingdom) demonstrates the approach by enabling Fault-Detection-and-Diagnosis of the Heating Ventilation and Air Conditioning system for facility management.

1. Introduction

Net-zero carbon objectives force restrictive goals and constraints towards the efficient use of energy in the built environment (Kazmi *et al.*, 2014; Ufuk Gökçe and Umut Gökçe, 2014). Efficient building operation remains as a big challenge in Architecture, Engineering, Construction and Facility Management (AEC/FM) where the focus is on the design and construction phases (Boje *et al.*, 2020). Building Information Modelling (BIM) became one of the main advancements and areas of research in AEC by enabling actionable and understandable 3D models of the built environment (Succar, 2009; Zhu *et al.*, 2023). Despite the promising research on BIM data generation for facility managers from early stages, there is a slow adoption of the technology in the industry where practitioners only implement maintenance strategies based on the as-built models received at handover (i.e., from construction phase to operations), and BIM is never updated (Azhar, 2011; Becerik-Gerber *et al.*, 2012; Volk *et al.*, 2014; Shigaki and Yashiro, 2021).

Building operations are characterised by systems functions (e.g., electricity, heating, plumbing, ...). In the current landscape, it is becoming popular to outsource operation of each function to third-party companies with the objective of effective equipment operation (Volk *et al.*, 2014), focusing on performance of individual systems by single-point or distributed monitoring (Kazmi *et al.*, 2014). However, these systems are not necessarily independent (e.g., HVAC needs from the electric system to operate, plumbing rely on mechanical system elements), and therefore, monitoring should be integrated. Some authors addressed the integration of Building Automation Systems (BAS) with BIM to improve single system's

operation, but complete integration of system functions has not been fully explored (Ufuk Gökçe and Umut Gökçe, 2014; Dong *et al.*, 2014; Oti *et al.*, 2016; Chen *et al.*, 2018a; Tang *et al.*, 2020; Quinn *et al.*, 2020; Hu *et al.*, 2021; Hosamo *et al.*, 2023a). All these improvements in individual system operations come at the cost of awareness and control for facility managers that need to rely on third-party service providers (Shigaki and Yashiro, 2021) which translates in the inability to orchestrate multiple building functions and limits efficient building operations.

Digital twin (DT) technologies are rising as the facilitators of integrated building operations in AEC/FM throughout the life-cycle of buildings (Dong *et al.*, 2014; Hu *et al.*, 2021, 2022; Khajavi *et al.*, 2019). In this domain, the Digital twins initiative is strongly influenced by BIM and it combines emerging technologies in the AEC/FM industry like Internet of Things (IoT) for environmental monitoring and resource tracking (Boje *et al.*, 2020; Sotres *et al.*, 2017). Nevertheless, digital twins are still hindered by the segregation of data generated by the delegation of building functions (e.g., HVAC, electricity, plumbing, ...) to third-parties. Outsourcing generates segregation of data since it is stored in siloes often controlled by third-party service providers (Hu *et al.*, 2016, 2021; Shigaki and Yashiro, 2021). This segregation is translated into data modelled to meet independent systems' requirements (e.g., building automation system, asset management system, occupancy, design and construction data) rather than as part of the overarching built environment entity (Corry *et al.*, 2015; Woodhead *et al.*, 2018). In addition, systems are distributed and buildings components information is often outdated, incomplete, and inaccurate when exchanged between the assets' life cycles (e.g., from design and construction to operations and management) (O'Donnell *et al.*,

2013). This paper is motivated by the need of adequate data integration and management as a crucial aspect for digital twinning of the built environment.

Some integration approaches in digital twins for the built environment include: BIM APIs, Extract-Transform-Load (ETL) processes, relational databases schema mapping, semantic web, and hybrid approaches. Among them, semantic web approaches (e.g., linked data and ontologies) became popular for built environment data integration in the last decade (Tang *et al.*, 2019; Kim *et al.*, 2018; Donkers *et al.*, 2022). Semantic web technologies drive data integration while enhancing understandability by achieving broad classification and description of built environment entities (Pauwels *et al.*, 2017; Corry *et al.*, 2015). Ontologies are domain specific (e.g., building structures and hierarchy, building functions structure, sensors), and, therefore, multiple ontologies are necessary to encompass all the intricacies and complexities of buildings (Terkaj *et al.*, 2017). The effort of creating an ontology that accommodates all domains (e.g., BIM, BAS, IoT, ...) becomes unmanageable as the digital twin escalates by incorporating more functions, systems, and digital twin applications. These ontologies are extended until they become hard to understand (Zhe *et al.*, 2006; Kumar and Baliyan, 2018; McDaniel and Storey, 2020; Hryhorovych, 2021). Some advancements in the domain of modularisation of semantic web ontologies show potential improvement of the reasoning capabilities by identifying core ontologies, and inter-connecting entities, but the approach lacks practical validation (Pauwels and Terkaj, 2016; Wagner *et al.*, 2022; Tan *et al.*, 2023). On top of that, high latency is introduced by ontology resolution which is a hindrance for real-time data provision (Neumann and Weikum, 2010; Bizer and Schultz, 2009; Eneyew *et al.*, 2022).

Data federation promotes the domain-specific independence of data sources while finding appropriate links to standardise integration of data (Van Der Lans, 2012b,a; O'Donnell *et al.*, 2013; Shen *et al.*, 2021; Barbella and Tortora, 2023). Data federation can use semantic web approaches combined with other technologies in order to avoid the burden of creating a combined ontology.

The objective of this paper is the integration of BIM (as-built information), IoT, and Building Automation Systems' data for real-time visualisations and applications in a digital twin environment to support facility management. The challenges faced are the connection of diverse domain (and often low-available) data and the real-time data provision. It is important to highlight that the data diversity challenge has been eased by translating domain data using industry-known ontologies like Industry Foundation Classes (IFC) (ISO, 2018), BrickSchema (Brick Consortium, Inc, 2023), and an adaptation of the Building Topology Ontology named ACP data model (Rasmussen *et al.*, 2021; Brazauskas *et al.*, 2021). The paper adopts a hybrid approach based on data federation and modularised semantic web also referred as federated data modelling throughout the paper. The technical aspects and methods for the integration of data in a digital twin in the built environment are the main focus of the paper. Techniques are demonstrated in a case study conducted on the digital twin of the Alan Reece building of the University of Cambridge. The case study demonstrates the integration methods used for a Fault Detection and Diagnosis (FDD) application of the building's Heating, Cooling and Air Conditioning (HVAC) system.

The rest of this paper is structured as follows. Section 2 explores the existing literature

related to digital twins in the built environment, information integration and the concept of data federation. Section 3 describes the methods and technological solutions used. Section 4 defines a case study where these methods were applied. Section 5 discusses caveats and limitations of this work. The conclusions derived from this research are drawn in section 6.

2. Related work

This section introduces the concept of digital twin in the context of the built environment and the relationship with BIM, BAS, and IoT. This relationship leads to the need for data integration. The concept of data federation is explained at the end of the section.

2.1 Digital twins in the built environment

A digital twin is a system that portraits the digital representation of a physical counterpart. Every digital twin needs to incorporate three perspectives: a digital representation (e.g., a 3D model), a flow of data from the physical world to the digital representation (e.g. monitoring), and a flow of data from the digital representation to the physical counterpart (e.g., actuation). Digital twins symbolise the natural convergence of emerging technologies in the AEC/FM industry like BIM, IoT, Artificial Intelligence (AI) towards integrated building functions monitoring and actuation. The roles of these technologies in digital twinning and their relationships are introduced in the subsequent paragraphs.

In this framework, BIM takes the role of the digital representation of the built-environment. BIM changed the way the built environment information is created, stored, and exchanged between involved stakeholders (Howell and Rezgui, 2018). The Industry

Foundation Classes (IFC) brought data exchange of BIM between industrial design applications (Autodesk Inc., 2021; Perttula and Suchocki, 2020; Building Smart Int., 2022; ISO, 2015). Model View Definitions (MVD) are subsets of the IFC that represent BIM data related to a specific discipline that support data exchange between BIM and Building Automation Systems.

The Internet of Things (IoT) and Building Automation Systems enable both monitoring and actuation for digital twins, pushing BIM towards adjacent research areas throughout the entire built environment life-cycle, at building, infrastructure, and city levels (Boje *et al.*, 2020; Sotres *et al.*, 2017; Angjeliu *et al.*, 2020). (Liebenberg and Jarke, 2023) proposes digital shadows (i.e., digital representation plus monitoring only) as accelerators for production engineering, operation, and service. (Hu *et al.*, 2022) identifies some challenges that Digital Twin designers face in the built environment including diverse and multi-function sensors systems, and multi-asset integration. (Hu *et al.*, 2022) classifies state-of-the-art Industry 4.0 technologies into Digital Twin solution areas like data acquisition, processing, modelling and simulation, and decision support enablers, and provides a technological framework to integrate all these technologies.

The role of the last element of digital twins for the built environment, Artificial Intelligence, is as part of the applications. Digital twin applications in the built environment include data visualisation (e.g., dashboards, heatmaps, navigation), condition monitoring (e.g., anomaly detection, Fault Detection and Diagnosis), and prognosis (e.g., fault prediction, predictive maintenance) (Hu *et al.*, 2022; Alanne and Sierla, 2022). Many are designed to

access and visualise integrated information dynamically in near-real time. Dynamic data collection may be active, via queries (e.g., RDBMS and SQL, LDAP, files sync), or passive, via subscriptions to data publishers (e.g., message-passing, MQTT, KAFKA, websockets, REST APIs, webhooks). Additionally, non-dynamic data sources like maintenance reports and schedules, assets specifications, and 3D models can support decision-making in digital twin applications. While dynamic access to these data sources is not critical, it facilitates digital twin applications development and non-dynamic data reuse. Latency becomes a requirement for the technologies in the digital twin framework used to enable end-user applications and actuation.

When relating these three disciplines, BIM, IoT, and BAS become data sources that enable digital twin applications. These applications define data and performance requirements for the integration methods that provide data. (Quinn *et al.*, 2020) highlights three challenges of linking BIM with live IoT data: integration methods, heterogeneity and availability of data, and suitability of data architecture to support static (e.g., semantic, geometric, and topographical) and dynamic (e.g., sensor data in the form of time-series) data. (Tang *et al.*, 2019) conducted an in-depth review of BIM and IoT devices integration in the AEC industry across multiple domains, including Facility Management. It highlights three key components of BIM and IoT integration:

- BIM serves as a data repository for contextual information. It consists of building geometry, and sometimes it is extended with IoT devices' description and location, static information, and other soft building information collected from occupancy patterns. Contextual information is stored in industrial BIM or IFC formats.

- Time-series data records sensor readings. Traditional time-series data is stored in well-structured relational databases. More recently, data is stored in time-series databases or No-SQL alternatives.
- The integration method between contextual information and time-series data. The review concludes with a classification of BIM and IoT integration methods: a) solutions using BIM tools' APIs and relational database, b) solutions transforming BIM data into a relational database using new schema, c) solutions creating new query language, d) solutions using semantic web technologies and e) hybrid approaches.

(Mohammed *et al.*, 2020) identifies that the 67% of the literature on IoT and BIM integration is framed during design and construction phases and only the 22% is contextualised within facility management. Among the latter, (Arslan *et al.*, 2017; Kirstein and Ruiz-Zafra, 2018; Wu *et al.*, 2018; Wang *et al.*, 2020; Kang *et al.*, 2018; Chen *et al.*, 2018b; Lu *et al.*, 2021; Hosamo *et al.*, 2023b) load BIM data into relational databases using new schemas (category b). Vendors enable the first two categories via APIs for information query from their proprietary BIM formats or model transformation into a relational database. There exist some open source tools that also enable such transformation from IFC (IFCOpenshell.org, 2021; Bock and Friedrich, 2023), but both options bring relational data modelling challenges that need to be addressed (Wyszomirski and Gotlib, 2020; Prudhomme *et al.*, 2020). Solutions creating new query language are limited in the context of digital twins since they do not enable dynamic query of sensor data (Tang *et al.*, 2019). Most research is focused on semantic web ontologies, and hybrid methods, described sections 2.2 and 2.3.

Accepted manuscript
doi: 10.1680/jsmic.23.00002

2.2 Data integration using semantic web approaches

Semantic web is the most popular technique used to store, share, and use heterogeneous data sets in the built environment (Quinn *et al.*, 2020). It uses the Resource Description Framework (RDF) to represent BIM, asset, and sensor information to act as a proxy to link them. (Tang *et al.*, 2019) enumerates the steps to data integration using semantic web:

- Transform contextual and reference data about spaces, assets, and sensors into RDF.
- Extract time-series data from its source and transform it into RDF.
- Link data silos across different domains via unique identification.
- Use query languages (e.g., SPARQL) to discover the relationships between all three components, including time-series data.

Table 1 summarises popular ontologies.

Recent developments try to combine IFC with semantic web technologies to facilitate data extraction. (Pauwels *et al.*, 2011) combined IFC and Express (ISO, 1994) into a knowledge base through RDF (W3C, 2014) to develop a semantic rule checking environment for the construction industry. IFCOWL exposes building structures and hierarchy, despite IFC being inherently biased towards 3D visualisation (Beetz *et al.*, 2009; Pauwels *et al.*, 2017; Ma and Liu, 2018). (Zhu *et al.*, 2023) extracts the semantics of IFC into a graph to improve the data access, query, and understandability. The Building Topology Ontology (BOT) (Rasmussen *et al.*, 2021) focuses on capturing topological concepts in buildings such as sites, floors, zones, and rooms. Some authors considered the representation of assets and systems in the built environment as part of BIM (Dave *et al.*, 2018; Tomasevic *et al.*, 2015) as the foundations of

digital twins in the built environment. Haystack (Haystack, 2021) is a popular ontology for describing building assets using semi-structured sets of tags which are highly custom but inconsistent. BrickSchema standardizes semantic descriptions of the physical, logical, and virtual assets in buildings and the relationships between them (Brick Consortium, Inc, 2023; Balaji *et al.*, 2016). Semantic Sensor Network (SSN) (W3C, 2016a) describes sensors and their observations, features of interest and samples, procedures, and actuators, and has been used to describe BAS data with semantic tags. The Smart Appliances Reference ontology (SAREF) (ETSI, 2023) is intended to enable interoperability between solutions from different providers and among various activity sectors in the Internet of Things (IoT). The Sensor, Observation, Sample, and Actuator (SOSA) ontology (Janowicz *et al.*, 2019; W3C, 2016b) redesigns SSN to provide lightweight general-purpose specification for modelling the interaction between the entities involved in the acts of observation, actuation, and sampling. Sensor Model Language (SensorML) (Open Geospatial Consortium, 2023) provides a robust and semantically-tied means of defining sensors, actuators, and processes associated with the measurement and post-measurement transformation of observations.

Many authors reviewed the semantic web literature to combine or reuse these ontologies to achieve data integration (Dibley *et al.*, 2012; Curry *et al.*, 2013; Costa and Madrazo, 2015; Zhang *et al.*, 2015; Terkaj *et al.*, 2017; Boje and Li, 2018; Boje *et al.*, 2020; Gouda Mohamed *et al.*, 2020; Hu *et al.*, 2021; Donkers *et al.*, 2022; Wang *et al.*, 2022; Zhu *et al.*, 2023; Eneyew *et al.*, 2022). The BACnet ontology (ASHRAE, 2013) reuses SSN, BOT, SOSA, and IFCOWL to describe spatial building data, assets, sensors, and values to enable better visualisation of the

automated systems (e.g., HVAC room diffusers) (Tang *et al.*, 2020).

(Quinn *et al.*, 2020) identifies some limitations in the use of ontologies for BIM, BAS, and IoT data integration. Most ontologies define custom extensions to create the link with BIM, and the need to manually map ontology tags to data points in the BAS remains a challenge (Chen *et al.*, 2018a). Additionally, the use of ontologies acting as a link proxy between BAS/IoT and BIM creates data redundancy and duplication, for instance, when converting time-series data into RDF. Because of the low performance of RDF representing fixed-structured data, the semantic web approach is both time consuming to implement and restricted to semantic data concepts represented in the ontology (Hu *et al.*, 2016). (Bradley *et al.*, 2016) identified a lack of information integration and governance despite BIM and the numerous semantic web approaches, and argue that these approaches suffer low scalability since they are geared towards knowledge discovery rather than dynamic integration and visualisation of data. (Bradley *et al.*, 2016) also identifies three underlying factors common to all integration approaches, namely, definition (i.e., vocabulary and metadata), process (i.e., alignment with operational aspects), and connection (i.e., dynamic association and use) of information. Despite the evolution of IFC into ontologies like IFCOWL, it has not fully solved the problem for all application domains since it from conception is designed towards visualisation rather than to be modified or used dynamically (Boje *et al.*, 2020). (Pauwels *et al.*, 2011, 2017) states that the limited expression range of the IFC causes limitations when one wants to describe a building using certain concepts not found in the IFC. IFC often contains multiple descriptions of the same information which creates difficulties in partitioning the

information (Terkaj *et al.*, 2017).

The latency introduced by ontology resolution is a challenge on its own that has been addressed from data engineering point of view (Neumann and Weikum, 2010). Ontologies are a way to enable data integration but most times at a cost of high latency (Bizer and Schultz, 2009). Query processors retrieve summarized time-series data using SPARQL, and its selection for time-series integration has large implications on the overall latency (i.e., time to provide an output) (Quinn *et al.*, 2020). For instance, (Chevallier *et al.*, 2020; Mavrokapnidis *et al.*, 2021; Donkers *et al.*, 2023) suggest query processors for integrating BIM and IoT data using linked data and known ontologies like SOSA, Brick, IFC and BOT. The impact of connecting real-time data through SPARQL queries in latency is higher than hybrid approaches, since this technology limits the benefits of the optimised relational and NoSQL data repositories, and message-passing protocols.

2.3 Data integration alternatives and hybrid approaches

Looking away from semantic web for BIM data integration, literature becomes scarce and disconnected, but focuses on leveraging BAS data. (Quinn *et al.*, 2020) relies on naming convention to extract summary time-series data from the IoT database, reducing data redundancy and facilitating implementation by avoiding the ontology mapping step. (Gerrish *et al.*, 2017) extracts BIM in Revit to JSON using Dynamo (Autodesk, 2023) and links it to BAS data in a relational database using a custom Python script. (Chamari *et al.*, 2022) suggests a hybrid approach to integrate IFC and BAS by combining triple stores and SPARQL for contextual data that connects assets and monitoring, and NoSQL databases for the time-series

data. Same authors repeat the method to integrate IoT and IFC information (Chamari *et al.*, 2023). (Cheng *et al.*, 2020) develops a data-driven predictive maintenance planning framework for Facility Management based on BIM, BAS, and IoT technologies, and achieves data integration by an ETL process into a SQL database. (Kang and Hong, 2015) designed an ETL process to integrate BIM, geographical data, and maintenance records in a custom DB. (Hosamo *et al.*, 2023a) integrates BIM, sensor time-series and maintenance records in a custom ETL process to enable condition monitoring. (Hadjidemetriou *et al.*, 2023) suggests an architecture for real-time and faster than real-time estimations that integrates geometric and systems data through simulation models and APIs.

Some hybrid approaches represent and store contextual and reference information through ontologies (building topology, sensor information, asset information), but retain sensor time-series data in relational databases (Eneyew *et al.*, 2022). Data integration is conducted via sensors' or assets' IDs described in RDF. (Corry *et al.*, 2015; Hu *et al.*, 2016) developed a hybrid architecture linking relational databases of time-series data through the sensor ID to sensor reference information using SSN and then used SPARQL to discover the relationships between sensors and semantically-described building contextual data in IFC. (McGlinn *et al.*, 2017) used a similar approach to store actuator and BIM data in RDF and used SPARQL to reason the inter-dependencies between them, then integrated the results with time-series data from sensors (including reference sensor data) from a relational database. (Tang *et al.*, 2019) states some benefits from hybrid approaches, including reduction of data duplicity of time-series data (RDF and relational database), storage saving since RDF format tends to become heavier, and better

query performance in both relational databases and RDF, respectively.

Data Lakes are considered a comprehensive approach for data management in distributed information systems (Kumar *et al.*, 2018). A data lake is a collection of storage instances of various data assets stored in a (near) exact copy of the source format (Gartner, 2023), and represent the natural evolution of data warehouses for distributed environments. Data pipelines are processes to prepare, clean, integrate, and provide access to data considering the individual information requirements of the data consumers and applications in a data lake (Mehmood *et al.*, 2019). Data lake architectures are a subset of Big Data architectures (ISO/IEC, 2020) where the focus is on enabling data ingestion from diverse and distributed sources while keeping data in its original raw state. Data is transformed on-demand using custom fit-for-purpose integration strategies that better suit data consumers and applications. There are plenty of data lake architecture examples (Fang, 2015; Ait Errami *et al.*, 2023; Mehmood *et al.*, 2019; Madera and Laurent, 2016; Chessell *et al.*, 2014, 2015), but generally they fit in the reference architecture in figure 1.

Looking at the limitations of non-semantic and hybrid approaches, data transformations and integration in the AEC context cannot be achieved effortlessly (Hu *et al.*, 2016; Adnan and Akbar, 2019). The lack of completeness and accuracy of the geometries and semantics is a common issue in ETL processes for construction data (Sani and Rahman, 2018). The need for fidelity (i.e., preserving raw data to avoid information loss) creates multiple versions of data, which induces a substantial risk of inconsistency (Sawadogo and Darmont, 2021). On-demand database schema mapping on large variety of sources is an arduous effort while integrating data

in the pipelines (Nargesian *et al.*, 2019). Data lakes can quickly turn into data swamps without appropriate management (Raj and Surianarayanan, 2020).

2.4 Data federation

Data federation can be considered as a hybrid approach for data integration. The idea of federation appeared with Big Data and HDFS (Hadoop, 2023). Federated data consist in ensuring independence and self-management of data sets instead of struggling to centralise them (Van Der Lans, 2012b). Techniques like data fusion and federated learning were developed to harness the value of data in such distributed environment.

Data fusion is a process of integration of multiple data representing the same real-world object into a consistent, accurate, and useful representation (Bleiholder and Naumann, 2009). Data fusion, in a more modern view of the concept, unlocks knowledge fusion across multiple disparate (but potentially connected) data sets and integrates the insights rather than schema mapping and data merging (Zheng, 2015). New-age data fusion methods can be stage-based (i.e., learning from each source at a time), feature level-based (i.e., extracting common features), and semantic meaning-based (i.e., extracting features, meaning and relationships). Federated learning, introduced in (Konečný *et al.*, 2016; McMahan *et al.*, 2017) and then popularised by Google, involves training models over remote devices or siloed data centres while keeping data localized (Li *et al.*, 2020). Federated learning could be regarded as a data fusion method where data remains at source and models are moved to the data sources and iteratively trained locally. Models' updates are sent back to a manager, decrypted, averaged, and consolidated into the centralized model. In the context of IoT, federated learning offers a way of

harnessing the potential of data streams from sensors when computing capabilities are moved to the edge (Chahoud *et al.*, 2023; Sun *et al.*, 2021). With the growing concern on data privacy at the edge, federated learning has been regarded as a promising solution for deploying distributed data processing and learning in wireless networks (Lu *et al.*, 2021).

In the case of digital twins, federation can be extrapolated to entire data sources in which data owners retain all data locally. Particularly, in the built environment data federation suits the natural segregation and independence of the data sources during operations and management (Hu *et al.*, 2016; Corry *et al.*, 2015). Each data source is modelled according to independent needs while collaborating for the integration towards specific Digital Twin applications. Federated approaches are necessary in this context to enable digital twin applications for data management (Moretti *et al.*, 2022), and for facility management (Qolomany *et al.*, 2020; White *et al.*, 2021; Pang *et al.*, 2021; Walters, 2019). (Werbrouck *et al.*, 2022) demonstrate how a Common Data Environments (CDEs) based on Linked Data for the AEC industry facilitate complex interactions between the various stakeholders participating in a project while maintaining independent federated data sets.

2.5 Gap analysis

Facility Management practice is characterised by the disconnection of the data which causes a lack of semantic interoperability. While semantic web solutions have demonstrated effective knowledge discovery, the need for manual tagging and the lack of performance of ontology resolution limit the performance of monitoring data stores in terms of real-time data delivery. In fact, some studies fell short to achieve real-time data integration due to the use of semantic

web approaches. Literature is moving towards adjacent areas by implementing hybrid approaches. Scalability has been explored in commercial architectures but mainly in terms of monitoring. One of the areas towards semantic interoperability that has not been fully explored in the context of digital twins for the built environment is data federation.

This paper targets real-time data integration for digital twins in the built environment using a federated approach. A data lake architecture is used to enable the connection to the federated data sources. The framework enables the use of different ontologies and custom data models to represent domain specific data. For instance, IFC is used to model BIM data (i.e., building components, geometries, and topology), BrickSchema is used to model the BAS (i.e., BAS components and hierarchy), and a custom flexible data model based in BOT is used for additional IoT sensor data and BAS operation records. The method for integration focuses on the sensor data stream and attaches the information required by the digital twin applications in a modular data pipeline, using transformations, schema-mapping, and semantic-based fusion to minimise latency. The method is demonstrated within a case study of a digital twin of a 3-storey building with several HVAC zones and a Fault Detection and Diagnosis (FDD) application.

3. Data integration for digital twins in the built environment

This section presents the architecture of the digital twin data platform and the method for BIM, BAS, and IoT data integration in the context of the built environment.

Accepted manuscript
doi: 10.1680/jsmic.23.00002

3.1 Digital twin data platform

The digital twin data platform is composed of services for data ingestion and storage, management, and consumption, shaping the architecture of a data lake. Well-known data cloud solutions for data platforms are commercially available, and other authors have suggested reference architectures for IoT-enabled smart buildings using such platforms focusing on scalability (Bashir *et al.*, 2022; Linder *et al.*, 2017, 2021; Genkin and McArthur, 2023). As part of a joint-research effort towards pushing commercial technology boundaries, the data platform selected is the Adaptive City Platform (ACP) (see figure 3) which was developed using state-of-the-art technologies like Vrtx to enable real-time applications with minimal end-to-end (i.e., from data sources to applications) latency (Brazauskas *et al.*, 2021). Data integration is enabled in this context, focusing on the real-time aspect of data.

The architecture is depicted in figure 2, and it has two flows of data: real-time streams and batch data.

The ACP ingests, stores, and manages data from real-time sources like IoT sensors (e.g., LoRaWan, radio-frequency, WiFi), and the BAS (e.g., HVAC components data points, including operating status and embedded sensors). It uses two forms of ingestion: publish-subscribe model (i.e., FeedHandler), and database connections (i.e., FeedMakers). The former is enabled by MQTT clients, and the latter by SQL clients that extract data with the required frequency. The ACP is engineered towards minimising the end-to-end latency for real-time data, averaging a few milliseconds between a data entry (i.e., when it is ingested) and exit (i.e., when it is available for use). This is particularly important in the built environment to

visualise the real status of assets and spaces and for the early identification of potential problems in Operation Maintenance and Repair (OM&R) (Boje *et al.*, 2020). Raw data is kept for traceability and repeatability. An internal high-end bus enables the internal flow of data using publish-subscribe model. Other services can subscribe to required data. Data storage of time-series data (i.e., MsgFiler) is redirected into day-level JSON files in the file system and, alternatively to a general purpose SQL database. The ACP enables access to data in real-time both through REST POST to the desired http destination URL (i.e., MsgRouteR), and by accepting data subscriptions through websockets (i.e., RTMonitor). More details about the ACP can be found in (Brazauskas *et al.*, 2021).

Other reference and transactional data sources may coexist in the data lake, but they are not necessarily ingested through the ACP. Reference data of the built environment consists of static blueprints, CAD drawings and 3D models of the building structures; documentation of mechanical, electrical, and plumbing systems, or other representations of their functional dependencies; and asset catalogues. Transactional data refers to semi-static information about status of assets, maintenance work orders, such as the condition inspection and date of assets. Reference and transactional data is stored either in its raw format or using a domain specific ontology like IFC for BIM or BrickSchema for the building functions. Access to these sources of data needs to be enabled by APIs (e.g., IFC Openshell, SPARQL, file system general I/O).

Both real-time streams and batch data are federated in multiple reservoirs in the data lake which facilitates the ingestion of high-variety data as well as high and low velocity data. Data pipelines enable pre-processing and integration.

3.2 Data pipelines and integration method

Data integration is enabled in this context, focusing on the real-time aspect of data. A similar approach is suggested by (Eneyew *et al.*, 2022), however, that method suffers from slower query response time for accessing a large number of semantically described sensor observations simultaneously. The relatively slower query response time is primarily the result of the query mediation overhead incurred during the serial transformation of the raw sensor data to an in-memory knowledge graph. This integration method circumvents that problem by extracting contextual information into memory in a hierarchical model that maps the BIM and BAS information, and then tags individual sensor readings realising them back to the real-time data stream.

Data integration is conducted through data pipelines. A data pipeline is a piece of software that sits between the data storage and sources, and the data consumers to extract, transform, and integrate available data on demand. The creation of data pipelines starts with the identification of information requirements by data consumers and applications (e.g., required data points, input format, pace) (Pishdad-Bozorgi *et al.*, 2018). A data pipeline can be reused by different applications and different data pipelines can be combined to serve a new application if required (see figure 4). It is essential to understand what data the data consumers and applications use, how often and what format it is required (Kang and Choi, 2015; Becerik-Gerber *et al.*, 2012). Sometimes, existing pipelines can be reused adding a new layer of data transformation (e.g., if a different format is needed) or integration (e.g., if additional data needs to be combined). If no existing pipeline can provide it, then a new one must be

designed from scratch. In such case, it is also necessary to identify the sources of data in the lake for the data pipeline to extract, transform and combine data. Ultimately, the data access is made available by data pipeline through an API in the format and pace required. In this paper, authentication is not considered but it is an important part of research in data lakes. This process is also known as data engineering.

The data integration method presented in this paper is focused on the real-time data stream from the ACP, but IoT or BAS data by itself lacks of meaning without context. Static sources add contextual information before applications make use of it (e.g., sensor location, building functions, subsystem). With the context, sensor data becomes a feature or an event that belongs to other entities. For instance, an IoT sensor produces temperature readings, but that temperature is the property of a space. Similarly, a vibration sensor in the BAS produces vibration frequencies in the X, Y, Z to represent the movement of a pump in the HVAC system.

IoT and BAS contextual data is modelled according to the ACP data model presented in (Brazauskas *et al.*, 2021), which is a schema governed by crates that structures data in a hierarchy similar to the BOT ontology (Rasmussen *et al.*, 2021). A crate is an entity (e.g., a sensor, an equipment, a space, a floor) with its own attributes plus zero or more parents (see figure 5). Parents are referenced in the crate approach rather than been nested. Thus, crates form a hierarchical structure, but every crate is uniquely identified through an indexed key for quick access. This is particularly effective to represent the topologies and functional hierarchies of buildings, their systems as well as sensor data. Documental databases are chosen to hold the crate model, and JavaScript Object Notation (JSON) format is used in the ACP.

Listing 1 shows an example of the crate model. Time-series data is collected through the ACP and every sensor reading is tagged with the contextual information based on the indexed id.

Even that the ACP data strategy is used to govern the flow of real-time data, there are two ontologies that help understanding sensor data in the built environment: the ISO 12006 (ISO, 2015) (Industry Foundation Classes - IFC data model) and BrickSchema.

IFC excels on 3D modelling (i.e., detail-drawing all the structures and components), and thus it has a complex representation the topologies and the architectural hierarchies, which can be inferred by traversing the IFC elements and relationships. (Moretti *et al.*, 2020) shows how IFC meta-information about hierarchy and topology of the built environment can be extracted. An IFC2ACP data pipeline was developed to read the IFC files and infer the topologies and architectural hierarchies using the IFCOpenShell python API (IFCOpenshell.org, 2021). Then it transforms that information into the ACP crates model and store it in memory in a JSON object to enable indexed access to all the elements. The IFC2ACP data pipeline reacts to changes in the original source on demand when a new IFC file is available.

The design of BrickSchema focuses on defining the physical, logical, and virtual building assets with the emphasis on building operations, such as equipment or sensors in lighting, sub-metering, and HVAC systems (Balaji *et al.*, 2016). BrickSchema is also effective for existing buildings where retrieving this information can be costly and time consuming (e.g., through laser scanning, imagery). Leveraging its strong expressability of building system hierarchies, the BrickSchema has been adopted in real cases like (Xie *et al.*, 2021) to represent building metering system, and to connect sub-metering readings with spatial characteristics for

fine-grained energy analysis. However, BrickSchema models are networks that contain many cycles to determine connections between entities. Traversing these cyclical networks in real-time can be costly in latency. (Xie *et al.*, 2021) shows how BrickSchema meta-information on the systems in the built environment can be transformed into the ACP crates data model for real-time applications. The BrickSchema files (i.e., turtle or TTL) are read in the Brick2ACP data pipeline using the py-brickschema python API (Brickschema.org, 2021), and transformed into the ACP crates data model to avoid that increase in latency. The ACP crates version of the BrickSchema meta-information is stored in JSON in memory for quick access. The Brick2ACP data pipeline reacts to changes in the original source on demand when a new TTL file is available.

The lingering question is how to integrate both ontologies since some applications may need to make use of metadata from both IFC and BrickSchema. Luckily, there is no need to create a meta-ontology to combine IFC and BrickSchema since they both handle the concept of a Space (in IFC) or Location (in BrickSchema) and Sensors (in IFC) or Points (in BrickSchema) which can be used as a nexus for the semantic mapping of the ontologies. A IFCxBrick data pipeline was implemented to integrate the tailored ACP versions IFC and BrickSchema data based on the common elements found in both ontologies. Transparent access to integrated data is enabled through an API. The API allows data consumers to access all elements in the building. It is also possible to query the elements inside the known element (i.e., children; e.g., all the sensors in a location) as well as the elements to which the known elements belongs to (i.e., the parents; e.g., the location of an asset/equipment).

Any application can open a websocket client connected to the ACP platform to subscribe to sensor data in real-time (see figure 6) and request contextual metadata from the IFCxBrick API.

Whereas semantic web approaches like centralised triple stores focus on data exploration, this approach is tweaked towards real-time data reporting. Developed APIs can also be used for exploration, but they are not specifically designed for data discovery. Despite having selected the ACP platform, this method could be also implemented using commercial tools that mirror the architecture and implement the pipelines. Integration in modular pipelines support the delegation of information management to data sources. Tagging individual readings with contextual data available in memory provide fast access to real-time data.

4. Case study: institute for manufacturing

The approach is implemented in the digital twin pilot of Alan Reece building at the West Cambridge site. The Alan Reece building is a 3-storey building and stands over a 3800-square-meter comprehensive area, including spaces for teaching, office, research, laboratory, canteen. Figure 7 shows the 3D model of the Alan Reece building. The digital twin is geared to support building operations and asset management. Among the applications enabled by the digital twin, this paper exemplifies the data integration approach through a Fault Detection and Diagnosis (FDD) functionality for building HVAC systems. The focus of this section is on the integration method application, rather than in the FDD application. Details on the FDD application can be found in (Xie *et al.*, 2021).

The setting for this case study is an HVAC zone comprising two seminar rooms where an

automated FDD application identifies anomalies in the comfort temperature of the spaces (see figure 8) using real-time analytics based on sensor and contextual data (i.e., IFC and BrickSchema). These two spaces are conditioned with a Variable Refrigerant Flow (VRF) system, connected to multi-zone indoor air conditioning units in a multi-split manner. Functionally, the VRF and indoor units provide heating and cooling the building, serving multiple seminar rooms and lecture theatres. The air conditioning system of the seminar rooms is pictured in figure 9. The FDD application makes use of temperature, humidity and dew-point, open-closed (for windows and doors) data from IoT sensors in the seminar rooms, plus operational data of the HVAC system from the BAS. Figure 9 shows the monitoring parameters in the BAS for this application.

For this example, the target is the Zone Temperature Malfunction fault (see figure 9). This fault refers to anomalies found on the comfort temperature of the seminar rooms' HVAC zone. If the temperature monitored by the IoT sensor exceeds the comfort interval, the FDD triggers an investigation. First, it checks the status of the windows and doors of the space, and the HVAC Zone temperature. If there is an anomaly, it is necessary to diagnose the sources, including potential impact to critical assets further up in the hierarchy of the HVAC system. Thus, operational status of the indoor units feeding the seminar rooms and the VRF needs to be checked accessing the BAS.

The information required for the FDD application consist of the real-time from the IoT sensors and BAS sensing points, the relationships between faults, assets, and spaces in the HVAC from BrickSchema, and the topology of the rooms from IFC. Real-time data is ingested

and managed by the Adaptive City Platform (ACP) which runs in a custom server. Reference data (i.e., IFC and BrickSchema models) is also stored in the same custom server in its original format. The IFC2ACP, Brick2ACP, and IFCxBrick data pipelines are responsible for the integration and online APIs are made available for all three data pipelines. The ACP enables realtime data requests and subscription through websockets.

Figure 10 shows the usage of the IFCxBrick data pipeline to discover and diagnose the zone temperature malfunction fault. The FDD application monitors the sensors related to the zone temperature malfunction fault.

The sequence starts with the FDD application querying the function getAllSensors indicating the fault id. The function will return the list of sensors related to this fault. Subsequently, the FDD application can subscribe to the real-time data of the sensor list. Then, on every message that arrives to the FDD the following steps may be triggered:

- The FDD application needs to know the type of sensor that the message is coming from. The type of sensor is in the body of the message coming from the ACP, but it can also be queried with the function getAllSensors of the IFCxBrick indicating the sensor id. In this example, IoT sensors are located only in spaces and never in equipment in the Alan Reece Building.
- The FDD application queries the function getAllSpaces of the IFCxBrick API with the sensor id to know what space (i.e., seminar room 2) that IoT sensor belongs to (i.e., its parent). If the temperature of the seminar 2 exceeds the comfort interval, the investigation is triggered.

- The FDD application will check other sensors in that space by querying the function getAllSensors of the IFCxBrick API with the list of spaces returned in the previous step. It also requests the last readings from the ACP through the websocket.
- A crosscheck including the original reading and the new sensor readings is performed to identify the source of the problem. All the sensors reporting a problem in the crosscheck must be investigated. The functions getAllEquipment and getAllSpaces from the IFCxBrick API can be queried again to know more about the HVAC system functional relationships (e.g., what HVAC Zone the space belongs to, what indoor unit feeds the HVAC Zone, which VRF feeds the corresponding indoor units; see figure 9), as well as the topology and the hierarchy of the building (e.g., what sensors are in a space, or an equipment, what spaces belong to other spaces; see figure 8).
- Similarly to step 3, the last sensor readings of the spaces and equipment that are subject to further investigation can be requested to the ACP. The last two steps are repeated until the source of the fault is identified.

The faults discovered by the FDD and the diagnosis can be reported through the required methods to assist facility management. In this case, the FDD reports via email to the facility manager, and the faults are visualised in the 3D model of the building by flashing the affected assets and spaces. Ideally, actuation can be enabled in the form of operating the HVAC system at a different temperature if there is no fault, but facility managers did not allowed full automation in this case study.

4.1 Evaluation

Comparison against the most similar methods in the literature is not completely possible. Technologies, hardware, and data is completely different in (Eneyew *et al.*, 2022), and, while their integration method focuses on knowledge discovery, the integration method suggested in this paper focuses on enabling real-time digital twin applications where it is key to minimise the latency introduced by the integration itself. The architecture suggested in (Hadjidemetriou *et al.*, 2023) is geared towards estimations and uses simulation models to integrate data, however, no metrics are provided on performance. Despite these differences, similar metrics are produced below.

This method attaches the contextual information to individual sensor readings using the pipelines. Data used for this evaluation was generated by twelve sensors over a period of three months. Sensors generated data between at different sampling rates over the same period: six of them generated data every two minutes, five of them every thirty seconds, and the last sensor generated data every second. For every single message passing through the pipelines, the latency introduced by the integration method was measured by taking a timestamp in nanoseconds before and after its execution. The average latency introduced by the method was 1.221 milliseconds with an standard deviation of 1,299 milliseconds, a maximum of 10 milliseconds and a minimum of 0,98 milliseconds. Since most reference data is loaded and indexed in memory, this method is really bounded to memory performance. As a reference, the integration method in (Eneyew *et al.*, 2022) is capable of integrating a single data point in 37 milliseconds.

Historical data queries metrics were produced using the same 3 months period. Figure 11 shows the integration time and the query response time over the number of sensor readings queried. The integration time is under the milliseconds barrier, however, this is diminished by the query time. In this case study, historical data is stored into JSON daily files in the file system which means that it needs to be loaded in memory from the hard-disk drives. This limits the potential of the architecture for this type of jobs. Historical data can be moved to a time-series data base to improve query times. Again, as a reference, (Eneyew *et al.*, 2022) queries (including integration) are around ten times faster, however, it does not provide exact metrics of how much the integration method adds to this query times. Authors assert that the larger part of the latency of queries was introduced by the SPARQL query resolution.

5. Discussion

Important lessons were learned during this case study. BAS are managed by separate roles and even departments, and serviced by different companies, reducing data availability. In most cases data is only available as spreadsheets or documents that require preprocessing (e.g., asset maintenance records, hand-over documents of buildings). Information changes are not documented, and original documents are still used as reference instead, affecting data accuracy and timeliness. It is easier to get a one-time dump of the information, but not on-demand. Even when on-demand access is enabled, the technical aspects of data engineering appear. The high-variety of BAS, the manifold applications' and users' requirements make the design of pipelines an ordeal. Further, the duplication of pipelines is likely without appropriate planning (e.g., shared features identified). Federated data models helped to manage this complexity

enabling the design of tailored models and modular data pipelines.

3D models of buildings are only available since the advent of BIM in the industry. Old built-environment documentation (e.g., floorplans, CAD) would need to be updated and digitalised to enable digital twins. Digital documentation of building functions like an HVAC system is even more unusual. The need of manually creating these digital documents is limited to skilled professionals that are knowledgeable in the domain and have been trained in the use of semantic web approaches, despite the existence of innovative tools like IFC modellers, BrickSchema and other ontology editors. This also applies to IoT deployment management and maintenance since sensors are often the forgotten assets in digital twins.

Digital twins should include actions triggered from the virtual environment to the physical counterpart. In the built-environment this perspective is often limited by the facility managers that still do not trust fully automated environments. Therefore, digital twins become decision support systems that can suggest appropriate actions. In the case of new building developments, digital twins designed from AEC stages have the potential of becoming multi-function systems to be directly handed over to facility managers.

6. Conclusions

AECO industry is steering in the direction of the digitalisation of buildings from the design and construction phase to ensure efficient building operations. Digital twins are enablers for smart buildings operations to meet Net-zero objectives, but they can only achieve their full potential when integrating building functions.

In this sense, one of the biggest challenges is to enable on-demand access to Building

Automation Systems (BAS) which ownership and management is often shared. Obtaining real on-demand data becomes an arduous cyclical process of requesting increasing access grants, since managers are not always willing to facilitate it. Senior asset managers must become facilitators in the development of the Digital Twin. The technical challenges of data engineering increase with high-variety data and uses, which may cause duplication of pipelines without adequate planning.

This paper demonstrates how to integrate available data from different building functions. BAS (e.g., the functional relationships of building systems and operational records), BIM (e.g., the topology, and the architectural hierarchy of buildings) and real-time IoT information are integrated to enable dynamic asset management applications that support efficient building operations towards the net-zero objectives (e.g., optimising the operation and maintenance of the HVAC system). The framework displays methods to connect diverse domain data using data federation, industry-known ontologies, and a data lake architecture. The framework unlocks the data stored in silos while the use of federated data models helps with the delegation of data responsibilities. The data pipeline design method presented illustrates how to elicit information requirements of asset management applications in order to identify original data sources and data transformations and combinations to meet them. The design of the pipelines based on data federation and real-time data streams enable real-timeliness in the case study, circumventing the limitations that other works suffer due to semantic web approaches to tackle interoperability. It shows how this integration methods can aid, not only condition monitoring, but also enhanced visualisation to promote facility management.

The insights and reports provided by the applications can serve as new sources of information, but it is necessary to understand how to re-purpose them to enable new analysis and insights. Federated data models may need to be extended to accommodate newly created asset management knowledge. Whereas data lakes support the design of digital twins for the built environment, data access to original sources will remain as a big challenge because of the chain of responsibility of data. Implementation of DTs in early stages of buildings life-cycle can help eliminate this burden, and can become the first step towards servitisation in operations.

Acknowledgements

This research forms part of the Centre for Digital Built Britain's (CDBB) work at the University of Cambridge within the Construction Innovation Hub (CIH). The Construction Innovation Hub is funded by UK Research and Innovation (UKRI), through the Industrial Strategy Fund.

This work was funded by UKRI grant NMZM/429. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

Adnan K and Akbar R (2019) Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management* **11**, 10.1177/1847979019890771.

Ait Errami S, Hajji H, Ait El Kadi K and Badir H (2023) Spatial big data architecture: From Data Warehouses and Data Lakes to the LakeHouse. *Journal of Parallel and Distributed Computing* **176**: 70–79, 10.1016/j.jpdc.2023.02.007.

Alanne K and Sierla S (2022) An overview of machine learning applications for smart buildings. *Sustainable Cities and Society* **76**: 103445, 10.1016/j.scs.2021.103445.

Angjeliu G, Coronelli D and Cardani G (2020) Development of the simulation model for Digital Twin applications in historical masonry buildings: The integration between numerical and experimental reality. *Computers & Structures* **238**: 106282, 10.1016/j.compstruc.2020.106282.

Arslan M, Riaz Z and Munawar S (2017) Building Information Modeling (BIM) Enabled Facilities Management Using Hadoop Architecture. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, IEEE, Portland, OR, pp. 1–7, 10.23919/PICMET.2017.8125462.

ASHRAE (2013) Building automation and control network ontology. https://bacowl.sourceforge.net/#.

Autodesk (2023) Dynamo BIM. https://dynamobim.org/.

Autodesk Inc. (2021) BIM Interoperability | openBIM &

buildingSMART | Autodesk. autodesk.com/industry/ aec/bim/interoperability.

Azhar S (2011) Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry. *Leadership and Management in Engineering* **11(3)**: 241–252, 10.1061/(ASCE)LM.1943-5630.0000127.

Balaji B, Bhattacharya A, Fierro G, Gao J, Gluck J, Hong D,

Johansen A, Koh J, Ploennigs J, Agarwal Y *et al.* (2016) Brick: Towards a unified metadata schema for buildings. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*, pp. 41–50.

Barbella M and Tortora G (2023) A semi-automatic data integration process of heterogeneous databases. *Pattern Recognition Letters* **166**: 134–142, 10.1016/j.patrec.2023.01.007.

Bashir MR, Gill AQ and Beydoun G (2022) A Reference Architecture for IoT-Enabled Smart Buildings. *SN Computer Science* **3(6)**: 493, 10.1007/s42979-022-01401-9.

Becerik-Gerber B, Jazizadeh F, Li N and Calis G (2012) Application Areas and Data Requirements for BIM-Enabled Facilities Management. *Journal of Construction Engineering and Management* **138(3)**: 431–442, 10.1061/(ASCE)CO.1943-7862 .0000433, publisher: American Society of Civil Engineers.

Beetz J, Leeuwen Jv and Vries Bd (2009) IfcOWL: A case of transforming EXPRESS schemas into ontologies. *AI EDAM* **23(1)**: 89–101, 10.1017/S0890060409000122.

Bizer C and Schultz A (2009) The Berlin SPARQL Benchmark. *International Journal on Semantic Web and Information Systems (IJSWIS)* **5(2)**: 1–24, 10.4018/jswis.2009040101.

Bleiholder J and Naumann F (2009) Data fusion. ACM Computing *Surveys* **41(1)**: 1:1–1:41,

10.1145/1456650.1456651.

Bock BS and Friedrich E (2023) IFCSQL. https://github.com/IfcSharp/IfcSQL.

Boje C, Guerriero A, Kubicki S and Rezgui Y (2020) Towards a semantic Construction Digital Twin: Directions for future research. *Automation in Construction* **114**, 10.1016/j.autcon.2020.103179.

Boje C and Li H (2018) Crowd simulation-based knowledge mining supporting building evacuation design. *Advanced Engineering Informatics* **37**: 103–118, 10.1016/j.aei.2018.05.002.

Bradley A, Li H, Lark R and Dunn S (2016) BIM for infrastructure: An overall review and constructor perspective. *Automation in Construction* **71**: 139–152, 10.1016/j.autcon .2016.08.019.

Brazauskas J, Verma R, Safronov V, Danish M, Merino J, Xie X, Lewis I and Mortier R (2021) Data management for building information modelling in a real-time adaptive city platform. *arXiv preprint arXiv:2103.04924* .

Brick Consortium, Inc (2023) BrickSchema. https://brickschema.org/ontology.

Brickschema.org (2021) Brick Ontology Python package. https: //github.com/BrickSchema/ py-brickschema.

Building Smart Int. (2022) buildingSMART - The International Home of BIM. www.buildingsmart.org.

Chahoud M, Otoum S and Mourad A (2023) On the feasibility of Federated Learning towards on-demand client deployment at the edge. *Information Processing & Management* **60(1)**:

103150, 10.1016/j.ipm.2022.103150.

Chamari L, Petrova E and Pauwels P (2022) A web-based approach to BMS, BIM and IoT integration: a case study. CLIMA 2022 conference 10.34641/clima.2022.228.

Chamari L, Petrova E and Pauwels P (2023) Extensible real-time data acquisition and management for IoT enabled smart buildings. 10.35490/EC3.2023.300.

Chen W, Chen K, Cheng JC, Wang Q and Gan VJ (2018a) BIM-based framework for automatic scheduling of facility maintenance work orders. *Automation in Construction* **91**: 15–30, 10.1016/j.autcon.2018.03.007.

Chen XS, Liu CC and Wu IC (2018b) A BIM-based visualization and warning system for fire rescue. *Advanced Engineering Informatics* **37**: 42–53, 10.1016/j.aei.2018.04.015.

Cheng JC, Chen W, Chen K and Wang Q (2020) Data-driven predictive maintenance planning framework for MEP components based on BIM and IoT using machine learning algorithms. *Automation in Construction* **112**: 103087, 10.1016/j.autcon.2020.103087.

Chessell M, Jones NL, Limburn J, Radley D and Shank K (2015) *Designing and Operating a Data Reservoir*. 1 edn., no. 1 in 1, Redbooks IBM, https://www.redbooks.ibm.com/abstracts/sg248274.html?Open.

Chessell M, Scheepers F, Nguyen N, van Kessel R and van der Starre R (2014) *Governing and Managing Big Data for Analytics and Decision Makers. Technical Report REDP-5120-00*, IBM, https://www.redbooks.ibm. com/abstracts/redp5120.html?Open.

Chevallier Z, Finance B and Boulakia BC (2020) A Reference Architecture for Smart Building Digital Twin. In *Proceedings of the International Workshop on Semantic Digital Twins,*

CEUR Workshop Proceedings, vol. 2615 (García-Castro R, Davies J, Antoniou G and Fortuna C, eds), CEUR, Heraklion, Greece, https://ceur-ws.org/Vol-2615/#paper2.

Corry E, Pauwels P, Hu S, Keane M and O'Donnell J (2015) A performance assessment ontology for the environmental and energy management of buildings. *Automation in Construction* **57**: 249–259, 10.1016/j.autcon.2015.05.002.

Costa G and Madrazo L (2015) Connecting building component catalogues with BIM models using semantic technologies: an application for precast concrete components. *Automation in Construction* **57**: 239–248, 10.1016/j.autcon.2015.05.007.

Curry E, O'Donnell J, Corry E, Hasan S, Keane M and O'Riain S (2013) Linking building data in the cloud: Integrating cross-domain building data using linked data. *Advanced Engineering Informatics* **27(2)**: 206–219, 10.1016/j.aei.2012.10.003.

Dave B, Buda A, Nurminen A and Främling K (2018) A framework for integrating BIM and IoT through open standards. *Automation in Construction* **95**: 35–45, 10.1016/j.autcon .2018.07.022.

Dibley M, Li H, Rezgui Y and Miles J (2012) An ontology framework for intelligent sensor-based building monitoring. *Automation in Construction* **28**: 1–14, 10.1016/j.autcon.2012.05.018.

Dong B, O'Neill Z and Li Z (2014) A BIM-enabled information infrastructure for building energy Fault Detection and *Diagnostics. Automation in Construction* **44**: 197–211, 10.1016/j.autcon.2014.04.007.

Donkers A, Yang D, de Vries B and Baken N (2022) Semantic Web Technologies for Indoor

Environmental Quality: A Review and Ontology Design. *Buildings* **12(10)**: 1522, 10.3390/buildings12101522.

Donkers A, Yang D, De Vries B and Baken N (2023) A Visual Support Tool for Decision-Making over Federated Building Information. In *Computer-Aided Architectural Design. INTERCONNECTIONS: Co-computing Beyond Boundaries*, vol. 1819 (Turrin M, Andriotis C and Rafee A, eds), Springer Nature Switzerland, Cham, pp. 485–500, 10.1007/978-3-031-37189-9_32.

Eneyew DD, Capretz MAM and Bitsuamlak GT (2022) Toward Smart-Building Digital Twins: BIM and IoT Data Integration. *IEEE Access* **10**: 130487–130506, 10.1109/ACCESS.2022.3229370.

ETSI (2023) SAREF: the Smart Applications REFerence ontology. https://saref.etsi.org/core/v3.1.1/.

Fang H (2015) Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, IEEE, Shenyang, China, pp. 820–824, 10.1109/CYBER.2015.7288049.

Gartner (2023) Gartner it glossary - data lake. https://www.gartner.com/en/information-technology/glossary/data-lake.

Genkin M and McArthur J (2023) B-SMART: A reference architecture for artificially intelligent autonomic smart buildings. *Engineering Applications of Artificial Intelligence* **121**: 106063, 10.1016/j.engappai.2023.106063.

Gerrish T, Ruikar K, Cook M, Johnson M, Phillip M and Lowry C (2017) BIM application to building energy performance visualisation and management: Challenges and potential. *Energy and Buildings* **144**: 218–228, 10.1016/j.enbuild.2017.03.032.

Gouda Mohamed A, Abdallah MR and Marzouk M (2020) BIM and semantic web-based maintenance information for existing buildings. *Automation in Construction* **116**: 103209, 10.1016/j.autcon.2020.103209.

Hadjidemetriou L, Stylianidis N, Englezos D, Papadopoulos P, Eliades D, Timotheou S, Polycarpou MM and Panayiotou C (2023) A Digital Twin Architecture for Real-Time and Offline High Granularity Analysis in Smart Buildings. *Sustainable Cities and Society* : 10479510.1016/j.scs.2023.104795.

Hadoop (2023) HDFS Architecture Guide. https://hadoop. apache.org/docs/r1.2.1/hdfs_ design.html.

Haystack (2021) Project Haystack. https://project-haystack.org/.

Hosamo HH, Nielsen HK, Kraniotis D, Svennevig PR and Svidt K (2023a) Digital Twin framework for automated fault source detection and prediction for comfort performance evaluation of existing non-residential Norwegian buildings. *Energy and Buildings* **281**: 112732, 10.1016/j.enbuild.2022.112732.

Hosamo HH, Nielsen HK, Kraniotis D, Svennevig PR and Svidt K (2023b) Improving building occupant comfort through a digital twin approach: A Bayesian network model and predictive maintenance method. *Energy and Buildings* **288**: 112992, 10.1016/j.enbuild.2023.112992.

Howell S and Rezgui Y (2018) *Beyond BIM: Knowledge management for a smarter built environment*. BRE Electronic Publications.

Hryhorovych V (2021) Construction of Normalized Metric for Hierarchical Data Structures based on Harmonic Functions. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, vol. 1, pp. 146–149, 10.1109/CSIT52700.2021.9648623, iSSN: 2766-3639.

Hu S, Corry E, Curry E, Turner WJN and O'Donnell J (2016) Building performance optimisation: A hybrid architecture for the integration of contextual information and time-series data. *Automation in Construction* **70**: 51–61, 10.1016/j.autcon.2016.05.018.

Hu S, Wang J, Hoare C, Li Y, Pauwels P and O'Donnell J (2021) Building energy performance assessment using linked data and cross-domain semantic reasoning. *Automation in Construction* **124**: 103580, 10.1016/j.autcon.2021.103580.

Hu W, Lim KYH and Cai Y (2022) Digital Twin and Industry 4.0 Enablers in Building and Construction: A Survey. *Buildings* **12(11)**: 2004, 10.3390/buildings12112004.

IFCOpenshell.org (2021) IfcOpenShell. http://ifcopenshell.org/.

ISO (1994) ISO 10303-11:1994 - Express. https://www.iso.org/standard/18348.html. ISO (2015) *ISO 12006-2:2015 Building construction — Organization of information about construction works Part 2: Framework for classification. Technical report*, International Organisation for Standardisation.

ISO (2018) ISO 16739-1:2018 - IFC. https://www.iso.org/standard/70303.html.

ISO/IEC (2020) *ISO/IEC 20547-3:2020. Technical International Standard 20547-3:2020*,

ISO/IEC, https://www.iso.org/standard/71277.html.

Janowicz K, Haller A, Cox SJ, Le Phuoc D and Lefrançois M (2019) SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics* **56**: 1–10, 10.1016/j.websem.2018.06.003.

Kang K, Lin J and Zhang J (2018) BIM- and IoT-based monitoring framework for building performance management. *Journal of Structural Integrity and Maintenance* **3(4)**: 254–261, 10.1080/24705314.2018.1536318.

Kang TW and Choi HS (2015) BIM perspective definition metadata for interworking facility management data. *Advanced Engineering Informatics* **29**: 958–970, 10.1016/j.aei.2015 .09.004.

Kang TW and Hong CH (2015) A study on software architecture for effective bim/gis-based facility management data integration. *Automation in Construction* **54**: 25–38, https: //doi.org/10.1016/j.autcon.2015.03.019.

Kazmi AH, O'grady MJ, Delaney DT, Ruzzelli AG and O'hare GMP (2014) A Review of Wireless-Sensor-Network-Enabled Building Energy Management Systems. *ACM Trans. Sen. Netw*. **10(4)**: 66:1–66:43, 10.1145/2532644.

Khajavi SH, Motlagh NH, Jaribion A, Werner LC and Holmstrom J (2019) Digital Twin: Vision, Benefts, Boundaries, and Creation for Buildings. *IEEE Access* **7**: 147406–147419, 10.1109/ACCESS.2019.2946515.

Kim K, Kim H, Kim W, Kim C, Kim J and Yu J (2018) Integration of ifc objects and facility management work information using Semantic Web. *Automation in Construction* **87**:

173–187, 10.1016/j.autcon.2017.12.019.

Kirstein P and Ruiz-Zafra A (2018) Use of Templates and The Handle for Large-Scale Provision of Security and IoT in the Built Environment. In *Living in the Internet of Things: Cybersecurity of the IoT - 2018*, Institution of Engineering and Technology, London, UK, pp. 29 (10 pp.)–29 (10 pp.), 10.1049/cp.2018.0029.

Konečný J, McMahan HB, Yu FX, Richtarik P, Suresh AT and Bacon D (2016) Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, p. NA, https://arxiv.org/abs/1610.05492.

Kumar S and Baliyan N (2018) Quality Evaluation of Ontologies. In *Semantic Web-Based Systems: Quality Assessment Models*, SpringerBriefs in Computer Science, Springer, Singapore, pp. 19–50, 10.1007/978-981-10-7700-5_2.

Kumar SAP, Madhumathi R, Chelliah PR, Tao L and Wang S (2018) A novel digital twin-centric approach for driver intention prediction and traffic congestion avoidance. *J Reliable Intell Environ* **4(4)**: 199–209, 10.1007/s40860-018-0069-y.

Li T, Sahu AK, Talwalkar A and Smith V (2020) Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* **37(3)**: 50–60, 10.1109/MSP.2020.2975749, conference Name: IEEE Signal Processing Magazine.

Liebenberg M and Jarke M (2023) Information systems engineering with Digital Shadows: Concept and use cases in the Internet of Production. *Information Systems* **114**: 102182, 10.1016/j.is.2023.102182.

Linder L, Montet F, Hennebert J and Bacher JP (2021) Big Building Data 2.0 - a Big Data

Platform for Smart Buildings. *Journal of Physics: Conference Series* **2042(1)**: 012016, 10.1088/1742-6596/2042/1/012016.

Linder L, Vionnet D, Bacher JP and Hennebert J (2017) Big Building Data - a Big Data Platform for Smart Buildings. *Energy Procedia* **122**: 589–594, 10.1016/j.egypro .2017.07.354.

Lu Y, Huang X, Zhang K, Maharjan S and Zhang Y (2021) Low-Latency Federated Learning and Blockchain for Edge Association in Digital Twin Empowered 6G Networks. *IEEE Transactions on Industrial Informatics* **17(7)**: 5098–5107, 10.1109/TII.2020.3017668.

Ma Z and Liu Z (2018) Ontology- and freeware-based platform for rapid development of BIM applications with reasoning support. *Automation in Construction* **90**: 1–8, 10.1016/j.autcon.2018.02.004.

Madera C and Laurent A (2016) The next information architecture evolution: the data lake wave. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, MEDES, Association for Computing Machinery, New York, NY, USA, pp. 174–180, 10.1145/3012071.3012077.

Mavrokapnidis D, Katsigarakis K, Pauwels P, Petrova E, Korolija I and Rovas D (2021) A linked-data paradigm for the integration of static and dynamic building data in digital twins. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ACM, Coimbra Portugal, pp. 369–372, 10.1145/3486611.3491125.

McDaniel M and Storey VC (2020) Evaluating Domain Ontologies: Clarification,

Classification, and Challenges. *ACM Computing Surveys* **52(4)**: 1–44, 10.1145/3329124.

McGlinn K, Yuce B, Wicaksono H, Howell S and Rezgui Y (2017) Usability evaluation of a web-based tool for supporting holistic building energy management. *Automation in Construction* **84**: 154–165, 10.1016/j.autcon.2017.08.033.

McMahan B, Moore E, Ramage D, Hampson S and Arcas BAy (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1273–1282, https:// proceedings.mlr.press/v54/mcmahan17a.html.

Mehmood H, Gilman E, Cortes M, Kostakos P, Byrne A, Valta K, Tekes S and Riekki J (2019) Implementing Big Data Lake for Heterogeneous Data Sources. In *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, IEEE, Macao, Macao, pp. 37–44, 10.1109/ICDEW.2019.00-37.

Mohammed BH, Safe N, Sallehuddin H and Hussain AHB (2020) Building Information Modelling (BIM) and the Internet-of-Things (IoT): A Systematic Mapping Study. *IEEE Access* **8**: 155171–155183, 10.1109/ACCESS.2020.3016919.

Moretti N, Xie X, Merino J, Brazauskas J and Parlikad AK (2020) An openBIM Approach to IoT Integration with Incomplete As-Built Data. *Applied Sciences* **10(22)**: 8287, 10.3390/app10228287.

Moretti N, Xie X, Merino Garcia J, Chang J and Kumar Parlikad A (2022) Developing a Federated Data Model for Built Environment Digital Twins. In *Computing in Civil*

Engineering 2021, American Society of Civil Engineers, Orlando, Florida, pp. 613–621, 10.1061/9780784483893.076.

Nargesian F, Zhu E, Miller RJ, Pu KQ and Arocena PC (2019) Data lake management: Challenges and opportunities. *Proc. VLDB Endow.* **12(12)**: 1986–1989, 10.14778/3352063.3352116.

Neumann T and Weikum G (2010) The RDF-3X engine for scalable management of RDF data. *The VLDB Journal* **19(1)**: 91–113, 10.1007/s00778-009-0165-y.

Open Geospatial Consortium (2023) Sensor Model Language (SensorML) | OGC. https://www.ogc.org/standards/sensorml.

Oti AH, Kurul E, Cheung F and Tah JHM (2016) A framework for information models for building design and operation. *Automation in Construction* **72**: 195–210, 10.1016/j.autcon.2016.08.043.

O'Donnell J, Corry E, Hasan S, Keane M and Curry E (2013) Building performance optimization using cross-domain scenario modeling, linked data, and complex event processing. *Building and Environment* **62**: 102–111, 10.1016/j.buildenv.2013.01.019.

Pang J, Huang Y, Xie Z, Li J and Cai Z (2021) Collaborative city digital twin for the COVID-19 pandemic: A federated learning solution. *Tsinghua Science and Technology* **26(5)**: 759–771, 10.26599/TST.2021.9010026.

Pauwels P and Terkaj W (2016) EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology. *Automation in Construction* **63**: 100–133, 10.1016/j.autcon.2015.12.003.

Pauwels P, Van Deursen D, Verstraeten R, De Roo J, De Meyer R, Van de Walle R and Van Campenhout J (2011) A semantic rule checking environment for building performance checking. *Automation in Construction* **20(5)**: 506–518, 10.1016/j.autcon.2010.11.017.

Pauwels P, Zhang S and Lee YC (2017) Semantic web technologies in AEC industry: A literature overview. *Automation in Construction* **73**: 145–165, 10.1016/j.autcon.2016 .10.003.

Perttula T and Suchocki M (2020) Development progress on IFC for infrastructure . Autodesk University. autodesk.com/autodesk-university/es/class/Development-Progress-IFC-Infrastructure-2020.

Pishdad-Bozorgi P, Gao X, Eastman C and Self AP (2018) Planning and developing facility management-enabled building information model (FM-enabled BIM). *Automation in Construction* **87**: 22–38, 10.1016/j.autcon.2017.12.004.

Prudhomme C, Homburg T, Ponciano JJ, Boochs F, Cruz C and Roxin AM (2020) Interpretation and automatic integration of geospatial data into the Semantic Web: Towards a process of automatic geospatial data interpretation, classification and integration using semantic technologies. *Computing* **102(2)**: 365–391, 10.1007/s00607-019-00701-y.

Qolomany B, Ahmad K, Al-Fuqaha A and Qadir J (2020) Particle Swarm Optimized Federated Learning For Industrial IoT and Smart City Services. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, IEEE, Taipei, Taiwan, pp. 1–6, 10.1109/GLOBECOM42002.2020.9322464.

Quinn C, Shabestari AZ, Misic T, Gilani S, Litoiu M and McArthur J (2020) Building automation system - BIM integration using a linked data structure. *Automation in Construction* **118**: 103257, 10.1016/j.autcon.2020.103257.

Raj P and Surianarayanan C (2020) Chapter Twelve - Digital twin: The industry use cases. In Advances in Computers, *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, vol. 117, Elsevier, pp. 285–320, 10.1016/bs.adcom.2019.09.006.

Rasmussen MH, Pauwels P, Lefrançois M and Schneider GF (2021) Building Topology Ontology. https://w3c-lbd-cg.github.io/bot/.

Sani M and Rahman A (2018) GIS and BIM integration at data level: A review. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 42, pp. 299–306, 10.5194/isprs-archives-XLII-4-W9-299-2018.

Sawadogo P and Darmont J (2021) On data lake architectures and metadata management. *Journal of Intelligent Information Systems* **56(1)**: 97–120, 10.1007/s10844-020-00608-7.

Shen W, Hu T, Zhang C and Ma S (2021) Secure sharing of big digital twin data for smart manufacturing based on blockchain. *Journal of Manufacturing Systems* **61**: 338–350, 10.1016/j.jmsy.2021.09.014.

Shigaki JSI and Yashiro T (2021) BIM and Automation of Building Operations in Japan: Observations on the State-of-the-Art in Research and Its Orientation. In *Proceedings of the 18th International Conference on Computing in Civil and Building Engineering* (Toledo Santos E and Scheer S, eds), Lecture Notes in Civil Engineering, Springer

International Publishing, Cham, pp. 879–894, 10.1007/978-3-030-51295-8_61.

Sotres P, Santana JR, Sánchez L, Lanza J and Muñoz L (2017) Practical Lessons From the Deployment and Management of a Smart City Internet-of-Things Infrastructure: The SmartSantander Testbed Case. *IEEE Access* **5**, 10.1109/ACCESS.2017.2723659, conference Name: IEEE Access.

Succar B (2009) Building information modelling framework: A research and delivery *foundation for industry stakeholders. Automation* i*n Construction* **18(3)**: 357–375, 10.1016/j.autcon.2008.10.003.

Sun W, Lei S, Wang L, Liu Z and Zhang Y (2021) Adaptive Federated Learning and Digital Twin for Industrial Internet of Things. *IEEE Transactions on Industrial Informatics* **17(8)**: 5605–5614, 10.1109/TII.2020.3034674.

Tan SZK, Kir H, Aevermann BD, Gillespie T, Harris N, Hawrylycz MJ, Jorstad NL, Lein ES, Matentzoglu N, Miller JA, Mollenkopf TS, Mungall CJ, Ray PL, Sanchez REA, Staats B, Vermillion J, Yadav A, Zhang Y, Scheuermann RH and Osumi-Sutherland D (2023) Brain Data Standards - A method for building data-driven cell-type ontologies. *Scientifc Data* **10(1)**: 50, 10.1038/s41597-022-01886-2.

Tang S, Shelden DR, Eastman CM, Pishdad-Bozorgi P and Gao X (2019) A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Automation in Construction* **101**: 127–139, 10.1016/j.autcon.2019.01.020.

Tang S, Shelden DR, Eastman CM, Pishdad-Bozorgi P and Gao X (2020) BIM assisted

Building Automation System information exchange using BACnet and IFC. *Automation in Construction* **110(November 2019)**: 103049, 10.1016/j.autcon.2019.103049.

Terkaj W, Schneider G and Pauwels P (2017) Reusing Domain Ontologies in Linked Building Data: The Case of Building Automation and Control. In *Proceedings of the Joint Ontology Workshops 2017*, CEUR-WS.org, p. online, https://www.semanticscholar.org/paper/Reusing-Domain-Ontologies-in-Linked-Building-Data%3A-Terkaj-Schneider/5533517181b3e78a52bb207d5ece00ee4cb0c562?sort=relevance&page=2.

Tomasevic NM, Batic MC, Blanes LM, Keane MM and Vranes S (2015) Ontology-based facility data model for energy management. *Advanced Engineering Informatics* **29(4)**: 971–984, 10.1016/j.aei.2015.09.003.

Ufuk Gökçe H and Umut Gökçe K (2014) Integrated System Platform for Energy Efficient Building Operations. *Journal of Computing in Civil Engineering* **28(6)**: 05014005, 10.1061/(ASCE)CP.1943-5487.0000288.

Van Der Lans RF (2012a) The Future of Data Virtualization. In *Data Virtualization for Business Intelligence Systems*, Elsevier, pp. 253–266, 10.1016/B978-0-12-394425-2.00013-7.

Van Der Lans RF (2012b) Introduction to Data Virtualization. In *Data Virtualization for Business Intelligence Systems*, Elsevier, pp. 1–26, 10.1016/B978-0-12-394425-2.00001-0.

Volk R, Stengel J and Schultmann F (2014) Building Information Modeling (BIM) for existing buildings — Literature review and future needs. *Automation in Construction* **38**: 109–127, 10.1016/j.autcon.2013.10.023.

W3C (2014) RDF - Semantic Web Standards. https://www.w3.org/RDF/.

W3C (2016a) Semantic Sensor Network Ontology. 10.5063/F11C1TTM.

W3C (2016b) Semantic Sensor Network Ontology (SOSA). 10.5063/F11C1TTM.

Wagner A, Sprenger W, Maurer C, Kuhn TE and Rüppel U (2022) Building product ontology: Core ontology for Linked Building Product Data. *Automation in Construction* **133**: 103927, 10.1016/j.autcon.2021.103927.

Walters A (2019) National Digital Twin Programme. https://www.cdbb.cam.ac.uk/what-we-did/national-digital-twin-programme.

Wang M, Altaf MS, Al-Hussein M and Ma Y (2020) Framework for an IoT-based shop floor material management system for panelized homebuilding. *International Journal of Construction Management* **20(2)**: 130–145, 10.1080/15623599.2018.1484554.

Wang T, Gan VJ, Hu D and Liu H (2022) Digital twin-enabled built environment sensing and monitoring through semantic enrichment of BIM with SensorML. *Automation in Construction* **144**: 104625, 10.1016/j.autcon.2022.104625.

Werbrouck J, Pauwels P, Beetz J and Mannens E (2022) Mapping Federated AEC Projects to Industry Standards Using Dynamic Views. In *Proceedings of the 10th Linked Data in Architecture and Construction Workshop co-located with 19th European Semantic Web Conference (ESWC 2022)*, CEUR, Hersonissos, Greece, pp. 65–76, https://ceur-ws.org/Vol-3213/.

White G, Zink A, Codecá L and Clarke S (2021) A digital twin smart city for citizen feedback. *Cities* **110**: 103064, 10.1016/j.cities.2020.103064.

Woodhead R, Stephenson P and Morrey D (2018) Digital construction: From point solutions to IoT ecosystem. *Automation in Construction* **93**: 35–46, 10.1016/j.autcon.2018.05.004.

Wu Cm, Liu Hl, Huang Lm, Lin Jf and Hsu Mw (2018) Integrating

BIM and IoT technology in environmental planning and protection of urban utility tunnel construction. In *2018 IEEE International Conference on Advanced Manufacturing (ICAM)*, IEEE, Yunlin, pp. 198–201, 10.1109/AMCON.2018.8615004.

Wyszomirski M and Gotlib D (2020) A Unified Database Solution to Process BIM and GIS Data. *Applied Sciences* **10(23)**: 8518, 10.3390/app10238518.

Xie X, Moretti N, Merino J and Chang JY (2021) Ontology-based spatial and system hierarchies federation for fine-grained building energy analysis. In *Proc. of the Conference CIB W78*, vol. 2021, pp. 11–15.

Zhang S, Boukamp F and Teizer J (2015) Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA). *Automation in Construction* **52**: 29–41, 10.1016/j.autcon.2015.02.005.

Zhe Y, Zhang D and YE C (2006) Evaluation metrics for ontology complexity and evolution analysis. In *2006 IEEE International Conference on e-Business Engineering (ICEBE'06)*, pp. 162–170, 10.1109/ICEBE.2006.48.

Zheng Y (2015) Methodologies for Cross-Domain Data Fusion: An Overview. *IEEE Transactions on Big Data* **1(1)**: 16–34, 10.1109/TBDATA.2015.2465959.

Zhu J, Wu P and Lei X (2023) IFC-graph for facilitating building information access and query. *Automation in Construction* **148**: 104778, 10.1016/j.autcon.2023.104778.

Accepted manuscript
doi: 10.1680/jsmic.23.00002

Listing 1. Crates model example for sensor data

```
"sensor-temperature-123456": {
  "acp id":
  "sensor-temperature-123456",
  "type": "sensor",
  "features": ["temperature"],
  "parents": [ { "parent id": "ifm-space-01",
      "parent type": "space"} ] },
"sensor-vibration-789012": {
  "acp id": "sensor-vibration-789012",
  "type": "sensor",
  "features": ["x", "y", "z"],
  "parents": [{ "parent id": "ifm-pump-01",
     "parent type": "equipment" } ] }
```

Accepted manuscript
doi: 10.1680/jsmic.23.00002

Listing 2. Example of the ACP representation of an IFCSpace

```
"103": {
  "crate id": "103",
  "crate type": "space",
  "acp ts": 1629813271.922424,
  "acp localtime": "2021-08-24T14:54:31.9224",
  "ifc id": "0UIH5Blo19ohldZ0jJVrWM",
  "ifc type": "IfcSpace",
  "parent crate id": "GF-basement",
  "ifc geometry": {
    "ifc geometry type": "IfcExtrudedAreaSolid",
    "ifc location": [71474.9958300488,29669.3674420791,-150.0],
    "ifc depth": 2375.0,
    "ifc sweptarea": {
      "type": "IfcArbitraryClosedProfileDef",
      "points": [
        [-2099.99980792596,-1137.49952716191], ...,
        [-2099.99980792596,-1137.49952716191]]
    }
  }
}
```

Accepted manuscript
doi: 10.1680/jsmic.23.00002

Listing 3. Example of the ACP representation of a BrickSchema location

```
"103": {
  "location_name":"http://ifm.cam.ac.uk/demo_building#103",
  "type": "https://brickschema.org/schema/Brick#Room",
  "number_of_points": 0,
  "number_of_equipment": 0,
  "parents": [{
     "parent_id": "http://ifm.cam.ac.uk/demo_building#Floor_1",
     "type": "https://brickschema.org/schema/Brick#isPartOf",
     "parent_type": "location"
    },{
     "parent_id": "http://ifm.cam.ac.uk/demo_building#Zone_103",
     "type": "https://brickschema.org/schema/Brick#isPartOf",
     "parent_type": "location"
    },{
      "parent_id": "http://ifm.cam.ac.uk/demo_building#LZone",
      "type": "https://brickschema.org/schema/Brick#isPartOf",
      "parent_type": "location"
    }]
}
```

Accepted manuscript
doi: 10.1680/jsmic.23.00002

Listing 4. Example of the integration of IFC and Brickschema through the ACP data model

```
"103": {
   "acp id": "103",
   "crate type": "space",
   "crate id": "103",
   "ifc": {
     "crate id": "103",
     "crate type": "space",
     "parent crate id": "GF-basement",
     ... see IFC listing ...,
   },
   "brick": {
     "location name": "http://ifm.cam.ac.uk/demo building#103",
     "type": "https://brickschema.org/schema/Brick#Room",
     ... see BrickSchema listing ...
   }
}
```

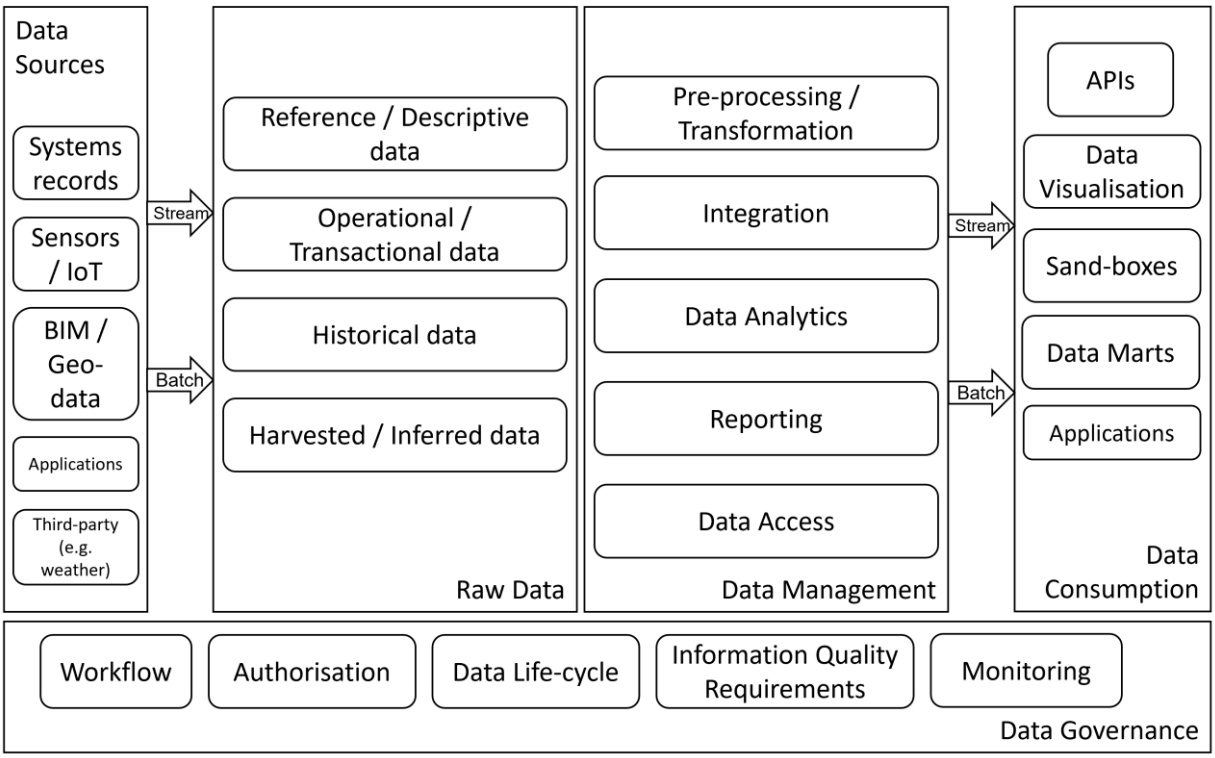
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Table 1. Semantic web ontologies for the built environment

Ontology	BIM	BAS	IoT
Industry Foundation lasses (IFC) (ISO, 2018) & IFCOWL (Beetz *et al.*, 2009)	*		
Building Topology Ontology (BOT) (Rasmussen *et al.*, 2021)	*		
Haystack (Haystack, 2021)	*	*	
BrickSchema (Brick Consortium, Inc, 2023)		*	
Semantic Sensor Network (SSN) (W3C, 2016a)		*	*
Smart Appliances Reference ontology (SAREF) (ETSI, 2023)		*	*
Sensor, Observation, Sample, and Actuator (SOSA) (W3C, 2016b)		*	*
Sensor Model Language (SensorML) (Open Geospatial Consortium, 2023)		*	*
Building Automation and Control Networks (BACnet) (ASHRAE, 2013)	*	*	*

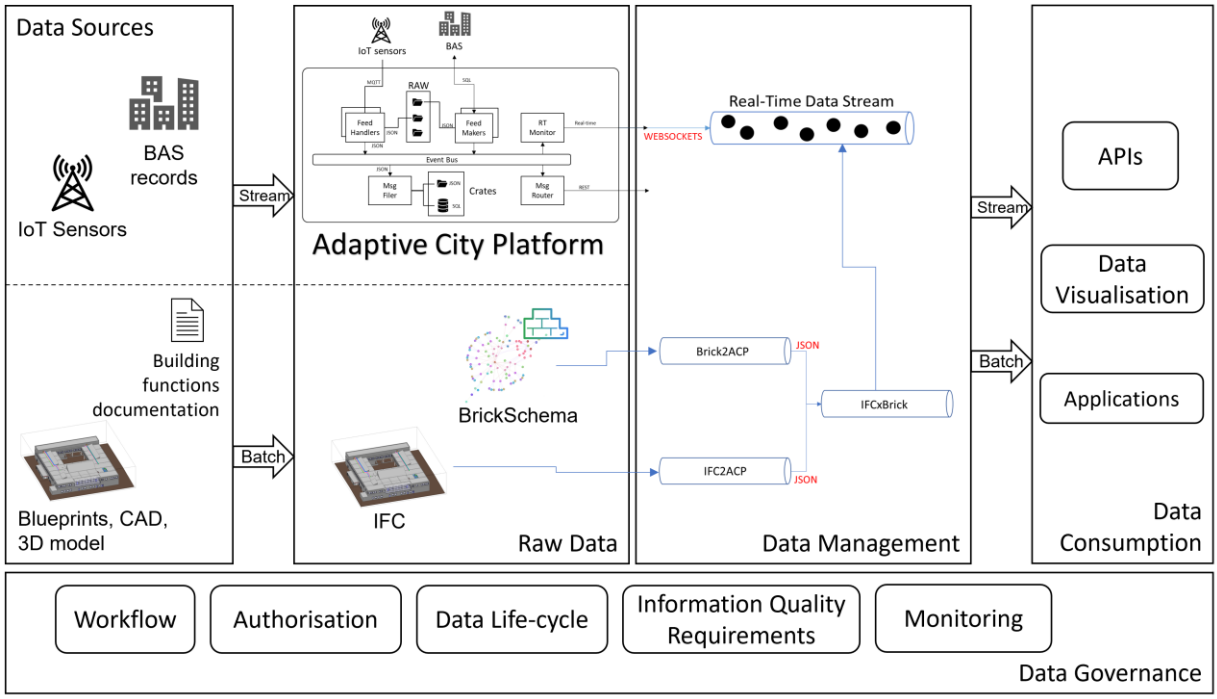
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 1. Reference Data Lake architecture



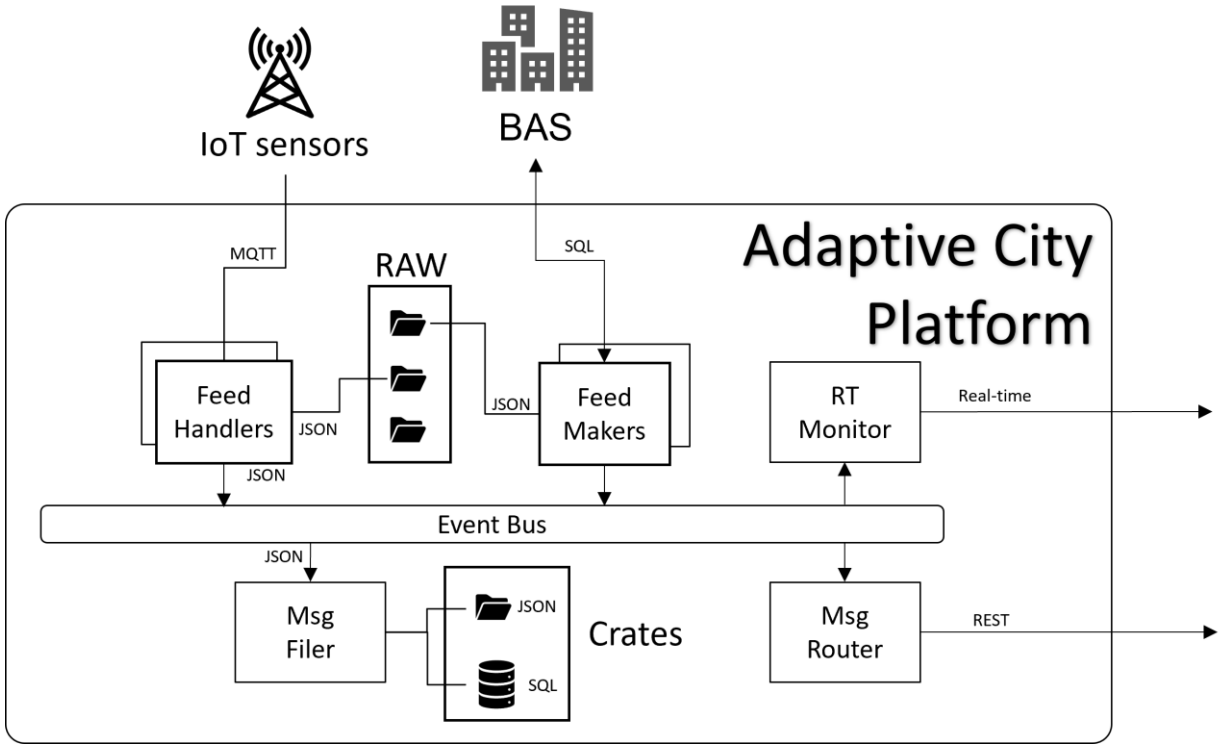
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 2. Digital Twin data platform architecture



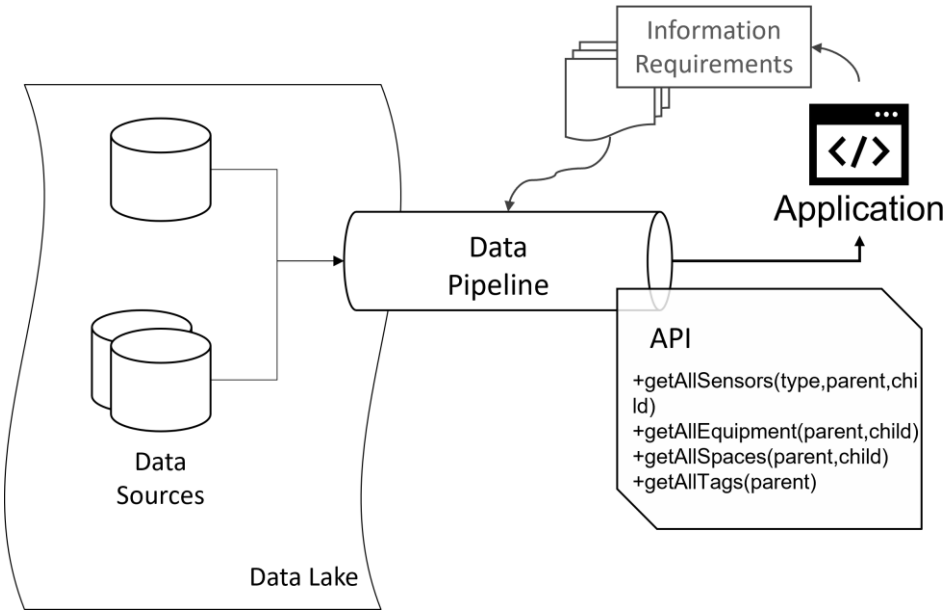
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 3. Adaptive City Platform (ACP) (Brazauskas *et al.*, 2021)



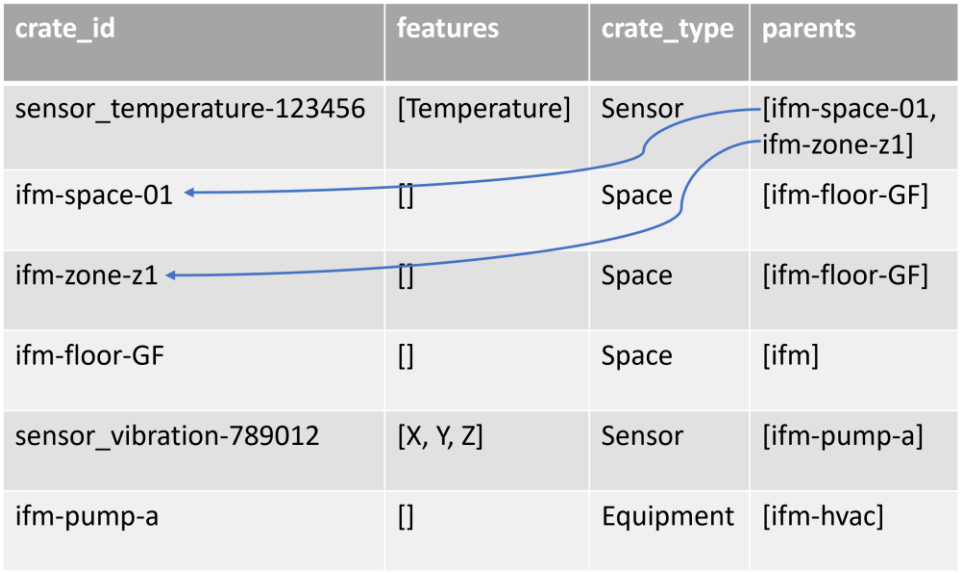
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 4. Data pipeline design: Inputs and outputs



Accepted manuscript
doi: 10.1680/jsmic.23.00002

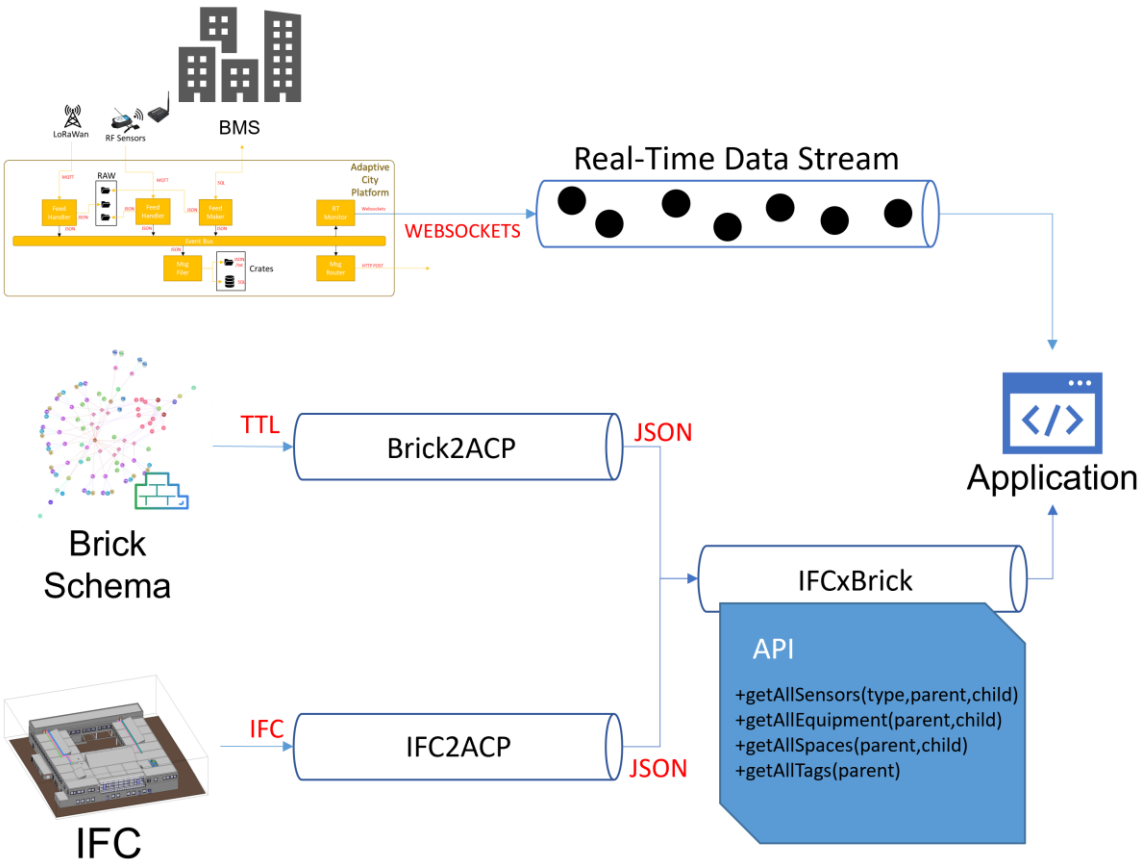
Figure 5. Adaptive City Platform (ACP) crates data model (Brazauskas *et al.*, 2021)



crate_id	features	crate_type	parents
sensor_temperature-123456	[Temperature]	Sensor	[ifm-space-01, ifm-zone-z1]
ifm-space-01	[]	Space	[ifm-floor-GF]
ifm-zone-z1	[]	Space	[ifm-floor-GF]
ifm-floor-GF	[]	Space	[ifm]
sensor_vibration-789012	[X, Y, Z]	Sensor	[ifm-pump-a]
ifm-pump-a	[]	Equipment	[ifm-hvac]

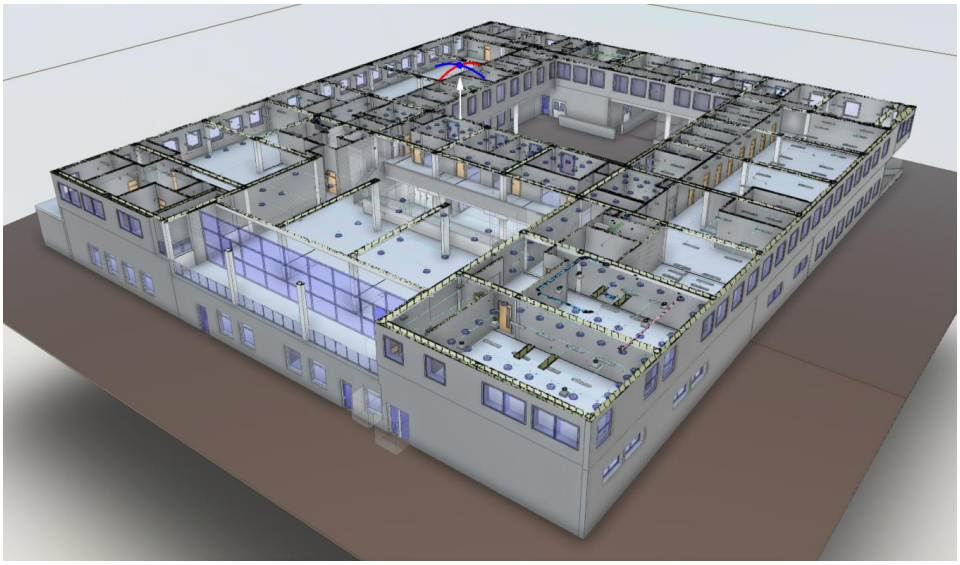
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 6. Data pipelines: ifc2acp, brick2acp and ifcxbrick, and the real-time connection with websockets



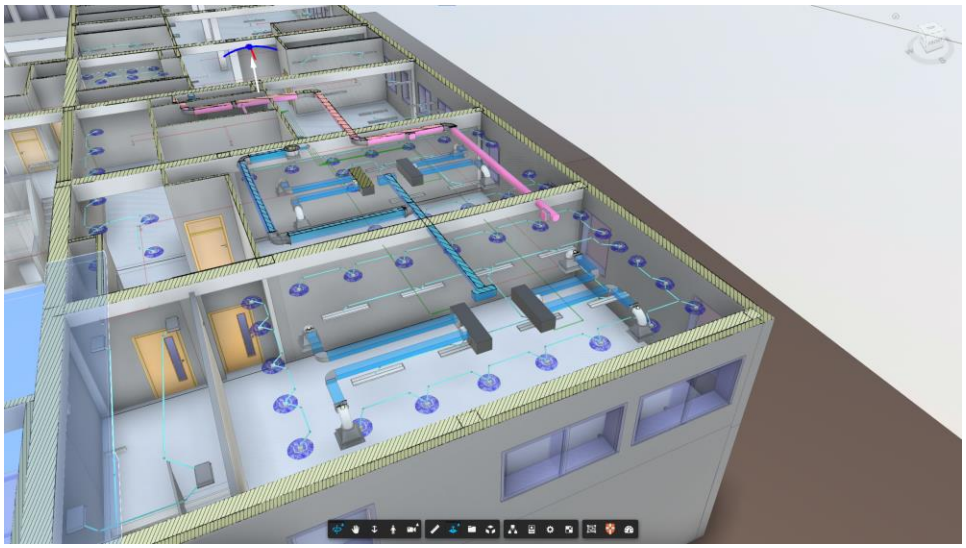
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 7. 3D Model of the Alan Reece building



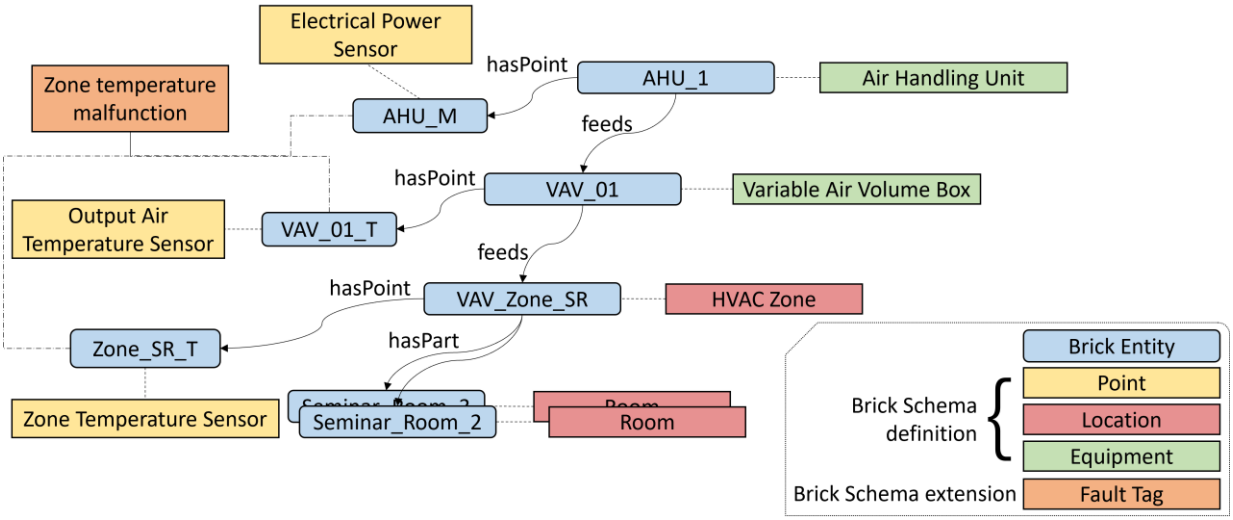
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 8. 3D model of the Seminar Rooms 2 and 3



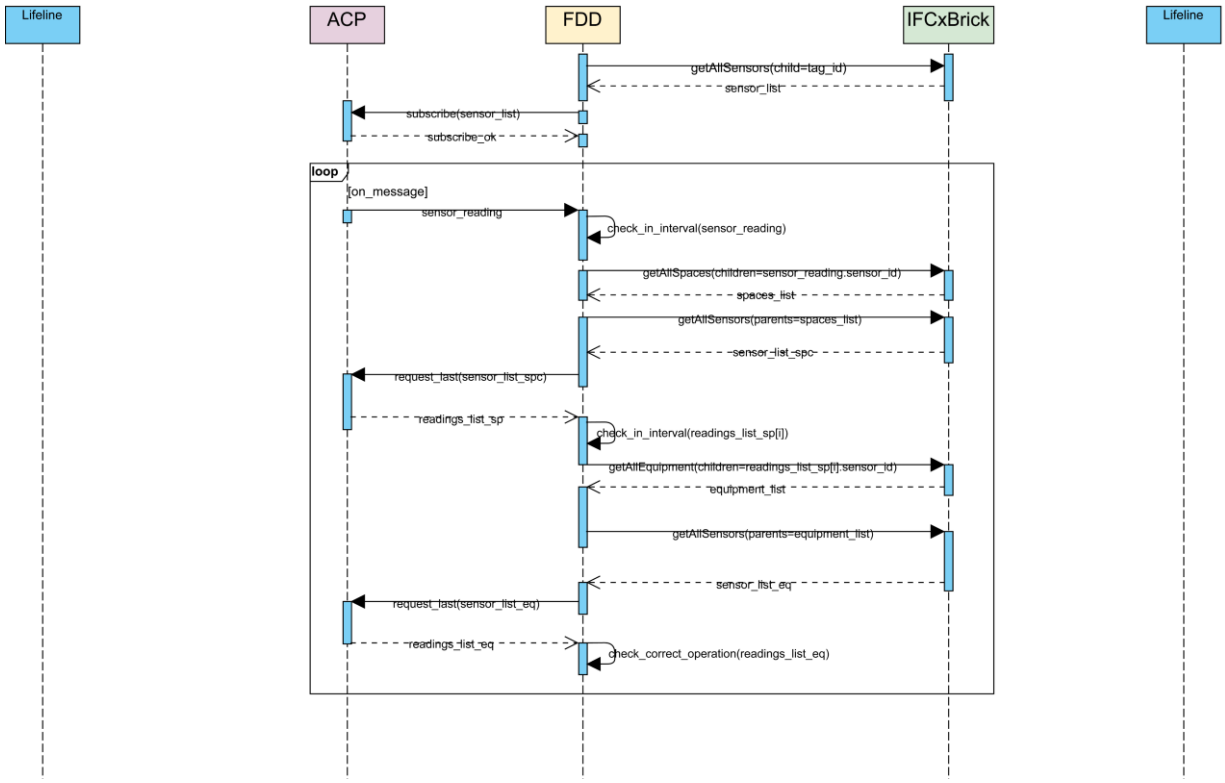
Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 9. BrickSchema model of the HVAC system feeding Seminar Rooms 2 and 3



Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 10. Sequence diagram to show the usage of the IFCxBrick API and the ACP websockets in an FDD application



Accepted manuscript
doi: 10.1680/jsmic.23.00002

Figure 11. Query response time evaluation