Knowledge Graph As Digital Twins Enhancer For Real Case Data-Driven Smart Building

Sebastien GERIN[1,2], Laurent JOBLOT[2], Nisrine MAKHOUL[1], and Frederic MERIENNE[2]

Abstract The integration of data capture, analysis, monitoring, and control technologies is rapidly becoming the cornerstone of next-generation smart buildings. However, developing digital twins that dynamically interact with these buildings presents a significant challenge. In this paper, we study the most appropriate data models for leveraging a digital twin from data-driven smart buildings. We proposed a framework that exploits a knowledge graph to directly address the challenges encountered in real-world building management systems, ensuring that the information is comprehensible as a preliminary step to intelligent decision-making. Furthermore, we validate this proposal for improving building performance and sustainability through a real-world use case. The experimental results, utilizing dynamic data streams from the Internet of Things (IoT), demonstrate promising outcomes. This research paves the way for using graph-based models and algorithms as digital twin enhancers for managing data-driven smart buildings.

Key words: Digital Twin; Data-driven Smart Buildings; Knowledge Graph, IoT

1 Introduction

The built environment significantly contributes to global CO2 emissions, reaching a record 10 Gt CO2 in 2021 [28]. This marks a 5% increase from 2020 and a 2% rise from the previous peak in 2019. Building operations, heavily reliant on fossil fuels for heating, cooling, and lighting, contribute up to one-third of total greenhouse gas emissions [28]. Various initiatives have been developed to mitigate this issue to improve building energy efficiency [10], like adopting renewable energy technologies such as solar panels, heat pumps, and energy storage solutions. These efforts are further enhanced by the emergence of "smart buildings" that leverage digital tools

---

[1]Institut de Recherche, ESTP, 28 Avenue du Président Wilson, F-94230, Cachan, France
[2]Arts et Métiers Institute of Technology, LISPEN / HESAM University / UBFC, Chalon-Sur-Saône, France

for real-time monitoring and optimization [12]. This paves the way for the digital twin (DT), an emerging concept increasingly utilized in design, manufacturing, and service industries [26]. This study focuses on finding the most appropriate method for creating a DT in the context of data-driven smart buildings (SB), where historical and real-time data are used to automate processes and decision-making [5]. To do so, an overview of data models for building management systems is conducted in Section 2. Section 3 presents a framework, integrating a knowledge graph and demonstrating its efficacy through an example of a practical application presented in Section 4. Section 5 suggests future research directions.

2 Related work

2.1 Preliminaries

DTs are real-time digital representations of a physical entity or process during its lifecycle. Updated in real-time with data from the real world, they allow precise reflection of any changes in a virtual environment. Initially used in aerospace, DTs now cover industrial manufacturing, healthcare, smart cities, education, or agriculture [18]. An ideal DT enables a two-way flow of information where real-time data updates the virtual representation for simulations and decision-making in the physical system [1]. This concept aligns with the idea of data-driven smart buildings (SBs) that utilize real-time and historical data to enable informed decision-making through data-drive processes [11, 29] to optimize energy use, indoor environment quality, and occupant experience [5]. The difference between a DT and Building Information Modeling (BIM) is a common point of confusion. BIM is an established paradigm for design and construction, aiming for trade collaboration, error reduction, or budget management. At the same time, DT focuses on real-time operational responses and optimization[18]. However, BIM enables the creation of detailed 3D building representations, which can be a fundamental foundation for creating a DT [7]. BIM can be integrated with Internet of Things (IoT) technology, a vast network of interconnected devices, to gather real-time data and achieve an ideal DT [25]. The process starts by using the 3D digital mock-up for design, construction, or operation, enabling simulation and analysis with the BIM model. Then, IoT integration can lead to energy performance analysis, space management, and more. Lastly, using AI algorithms for automated predictions, such as energy optimization, performs an ideal DT for autonomous buildings that provide feedback to decision support tools during its lifecycle [8].

2.2 Data models in data-driven smart building

In the literature, some authors point out the absence of standardization and ready-made solutions in creating a digital representation of buildings, which is essential for making informed decisions in digital transformation. One limitation was the lack of standards for connected building data models [5]. However, few core metadata schemas have been proposed and compared in the literature [2, 11, 20]. We propose an overview and a classification of the core metadata models covering most building concepts. Core metadata schemas adhere to the same semantic graph-based data model paradigm, which emphasizes the concept of knowledge graphs (KGs), crucial for data contextualization and exchange, information integration, and knowledge sharing [23]. A KG can be defined as a multi-relational graph comprising entities (nodes or vertices) and relations (edges). Each edge is represented as a semantic triplet, also known as a fact. A KG is ideal for enhancing a DT by managing heterogeneous and dynamic data [24]. The identified KG models for SB can be divided into two different and partly incompatible paradigms: *RDF models* and *labeled property graphs (LPGs)* [2]. RDF graph data models [15] originated from standard technology provided by the semantic web, structure information as a 'subject-predicate-object' triplet, representing two nodes linked by a single edge. Semantic web technologies can aid in flexible data integration across different domains, promoting interoperability among data and systems for DT [27]. This paradigm leverages ontologies, formal representations of domain concepts and relationships, for standardized logical reasoning [13], as part of deductive knowledge [17]. Ontologies rely on specific languages like Ontology Web Language (OWL) to ensure consistency and data quality with inference and constraint checking. RDF-based knowledge graphs are stored in a purpose-built triple store database, which can manage and retrieve triples through semantic queries. Building Ontology Topology (BOT)[1] and SSN/SOSA[2] are notable ontologies examples which capture building structures and sensor observations, respectively [21, 6]. Brick[3], maintained by the Brick Consortium since 2021, uses OWL to standardize semantic descriptions of physical, logical, and virtual building elements and their connections [3]. SAREF4BLDG[4], an extension of the SAREF ontology (for Smart Application Reference) is designed for buildings by incorporating Industry Foundation Classes (IFC) standard [19]. In another way, LPG consists of nodes and edges, each with its properties, which means that both nodes and edges can contain data attributes. LPG utilizes a NoSQL graph database, providing high performance for applications requiring data interconnectivity [9]. LPG is unsuitable for reasoning but can use inductive knowledge to make predictions based on generalized patterns from data observations. This approach can use Graph Machine Learning (GML), which applies machine learning algorithms to graph data for tasks like node property prediction (supervised) or community detec-

[1] https://w3c-lbd-cg.github.io/bot/

[2] https://github.com/w3c/sdw-sosa-ssn

[3] https://brickschema.org/ontology/

[4] https://saref.etsi.org/saref4bldg/v1.1.2/

tion (unsupervised). (REC)[5] is a modular model developed in 2017 by a community of developers (industry, consultancy, etc.); it describes concepts and relations in building data [14]. Initially designed in OWL, REC has been rewritten in Digital Twin Definition Language (DTDL) [6], adopting the LPG format [7]. Google Building Digital Ontology (DTO) is another data schema developed by Google in 2019 to represent structured building [4]. DTO is defined in YAML format to adopt the LPG paradigm.

3 The proposed approach

It has been established in the literature that no single data model captures all necessary smart building characteristics, making it impossible to determine one's superiority over another. The optimal choice depends on the purpose and the specific use case considerations [20, 11]. Concerning the graph-based data model format, LPG outperforms RDF in all metrics, although RDF's standardized format allows for better interoperability and domain knowledge sharing among stakeholders [2]. This section provides an overview of the requirements for selecting the most appropriate data model for a true data-driven smart building.

3.1 Pre study and model picking

This study is conducted at the *Dijon Metropolitan Campus*, a 10,000-square-meter smart building with the Ready2Service (R2S) certification for its digital services. The campus environment includes occupancy, temperature, CO2 levels, and energy consumption sensors, generating over 12,000 metadata entries and up to 1.7 million daily updates stored in a schema-less MongoDB database in JSON format. Given the scale of this 'big data,' a data-driven approach is preferable for managing and sharing data efficiently. Utilizing an open standard, defined and recommended by the W3C, ensures interoperability, reliability, and quality assurance. Finally, as a research hypothesis for future work, we aim to explore the potential of graph machine learning (GML) in analyzing the vast quantities of graph-based structured data available to us. At this stage, we believe the REC ontology is the most suitable for our needs at the *Dijon Metropolitan Campus*. REC employs the LPG graph format as a data-centric model, capable of performing real-time operations to store and access vast amounts of data through graph databases. Although the principal limitation of LPG is the absence of standardization, this gap is addressed using DTDL, a specification language based on JSON-LD, which is an open format recommended

[5] https://github.com/RealEstateCore/rec
[6] https://learn.microsoft.com/en-us/azure/digital-twins/concepts-models
[7] https://dev.realestatecore.io/docs/DTDL-or-SHACL

by W3C. JSON-LD is well-suited to the existing databases of industrial partners, which store data in JSON format, eliminating the need for an additional database to store the transformed data. In their review, Tang and Al discussed five integration methods between BIM and IOT, from combining BIM with a relational database or semantics web approach using existing RDF-based ontologies to hybrid ones mixing relational and semantic [25]. We believe that using a property graph as a semantic model, defined with JSON-LD, for converging BIM topology and time-series data values represents a novel approach complementing the previously mentioned methods. This work can be seen as an initial step in evaluating the most suitable model for the campus.

3.2 *Workflow*

This section presents a workflow to address the standardization gap for DT in SB, as illustrated in Figure 1. The pipeline operates in three key stages: *(i) data collection*, *(ii) ETL algorithm*, and *(iii) visualization*. In the following, we will detail each step.

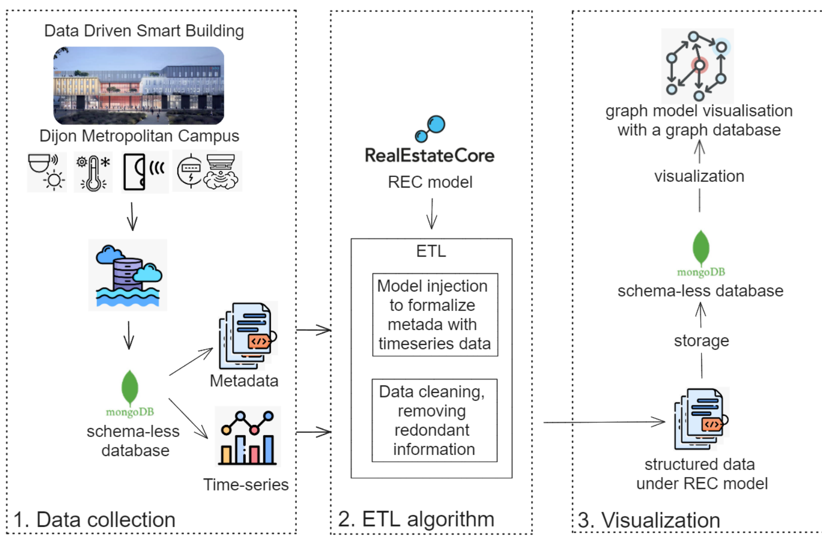

Fig. 1 Architecture of the proposed framework.

3.2.1 Data collection

Buildings possess numerous heterogeneous data sources that are often complex to grasp and access. They can include a list of IoT identifiers, time series data from systems, and BIM models that can be stored in different databases. The metadata associated with each IoT point can refer to device type, control command, and building topology, including properties relative to standards or defined explicitly by the vendor. The Building Operating System (BOS) is the primary aggregator of this data in SB. It standardizes diverse data from interconnected devices and systems, facilitating compatibility with third-party applications.

3.2.2 ETL algorithm

The ETL[8] process aims to extract and transform metadata and time-series data into structured data following the REC semantic model and format. The process is described as follows: The algorithm takes as input the metadata M, time-series data T from the data-driven smart building, and the extraction dates t_1 and t_2. The algorithm's output is structured data S conforming to the REC semantic model.

Algorithm 1: The ETL algorithm for REC model transformation

Data:

- Metadata M, time-series data T, t_1, t_2.

Result:

- Structured data with REC model S.

```
  begin
1     foreach ( id ∈ M ) do
2         if (CONTAINS_TOPOLOGY(M, E) ) then
3             UPDATE_MODEL(M, E, S) ;
4             GET_VALUES(id, T, t1, t2, v, t) ;
5             CLEAN_DATA(id, v, t);
6             BUILD_SEMANTIC_MODEL(S, t,v) ;
7 return S;
```

We extract all the entry points from the metadata collection M (Line 1). The topology is checked at each entry point using the Contains_Topology function, which validates and identifies building topology elements. The Update_Model procedure extracts site, zones, floors, and space instances from the building metadata and creates the corresponding REC topology JSON-LD node from the established model. In the same way, Get_Values retrieves values and timestamps from the time series and creates the corresponding REC node. It aims to enrich time-series data with standard semantics. Clean_Data eliminates the contiguous redundant values for optimization. The Build_Semantic_Model procedure enriches the model by creating relationships between the extracted topology node, the sensor node, and the cleaned time-series values nodes formatted in the REC model in JSON-LD format.

[8] The ETL acronym stands for extract, load, and transform.

3.3 Visualization

The final step of our framework involves loading the exported REC graph schema into a graph database management system for visualization and manual verification of the extracted LPG. We utilize Neo4J, the market leader and most widely used graph data platform, to facilitate this process [22].

4 Experimental results

The results obtained by applying our proposed approach will be presented, focusing on the occupancy sensor for a given day. The Linksper BOS [9] at the *Dijon Metropolitan Campus* uses a tag system to describe entry points and the building's topology. For each occupancy entry point [10], the Update_Model and Build_Semantic_Model procedures create room node [11], which owns a *has part* relationship to an occupancy sensor [12], respecting the REC format. Similarly, nodes containing boolean observation values are created and linked to the REC semantic model through a boolean observation value node [13]. These nodes correspond to the time series for a specific day, which consists of a series of boolean event values along with timestamps [14]. A value of True indicates that a person has entered the room, while a value of False signifies that the room is empty. The Clean_Data method has been updated to eliminate redundant values and false readings for boolean occupancy sensor observations within a 30-minute interval. This interval accounts for situations where an occupant is present but not moving enough to trigger the sensor. This delay is arbitrary and will need to be adjusted in future work. For a specific day, the ETL algorithm retrieves occupancy sensor data. It generates 1,980 JSON-LD documents in under 20 seconds using a laptop computer with an Intel Core i7 processor and 32 GB RAM. Without time-series cleaning, 14,640 JSON-LD documents are exported.

4.1 ETL Results

Figure 2 depicts a part of the resulting KG, where pink node represents the Campus site, and yellow node the building itself. Light blue nodes denote various floors within each building, linked to corresponding room spaces represented as red nodes.

---

[9] https://vayandata.com/bos-tertiaire/

[10] https://github.com/sebGrn/dijonMetropolitanCampusData/blob/main/metadata.json

[11] https://github.com/sebGrn/dijonMetropolitanCampusData/blob/main/rec_room_node.json

[12] https://github.com/sebGrn/dijonMetropolitanCampusData/blob/main/rec_occupancy_sensor.json

[13] https://github.com/sebGrn/dijonMetropolitanCampusData/blob/main/rec_boolean_value_observation.json

[14] https://github.com/sebGrn/dijonMetropolitanCampusData/blob/main/timeserie.json

Dark blue nodes illustrate occupancy sensors connected to each room, while green nodes designate boolean occupancy values with associated timestamp properties for each sensor. Figure 3 shows an example of a boolean value observation, with value and timestamp properties representing the room occupancy.

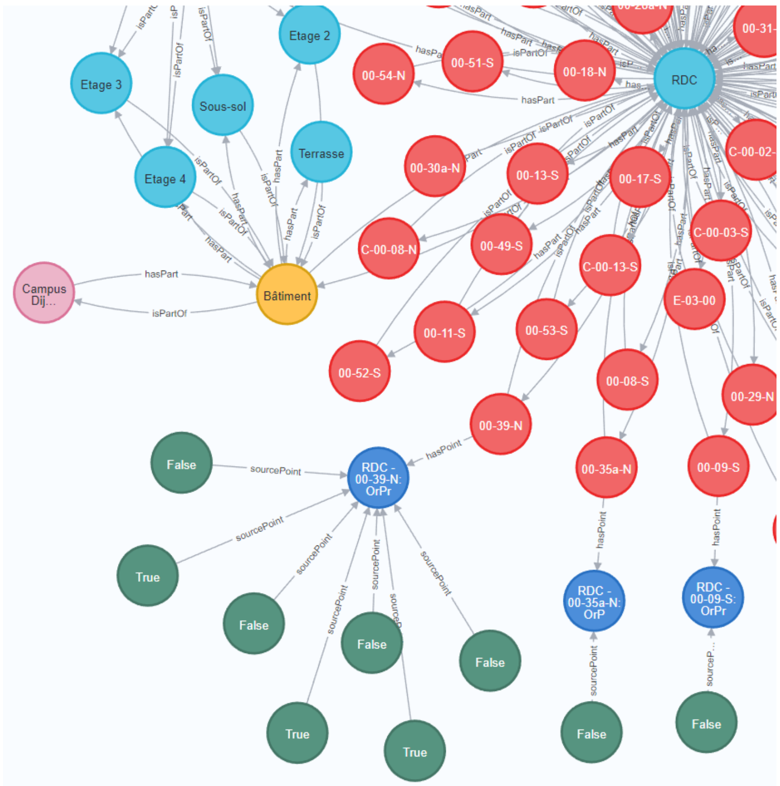

Fig. 2 Knowledge graph encompassing the various instances of the Dijon Metropolitan Campus encapsulated within the RealEstateCore model.

4.2 Graph queries

In this section, After organizing the campus building metadata and time-series data within a structured graph-based schema based on REC, we describe the relevance of graph algorithms and leverage semantic relationships in a series of Cypher[15]queries, the declarative query language designed for Neo4j. We propose illustrating the ini-

[15] https://neo4j.com/docs/cypher-manual/current/introduction/

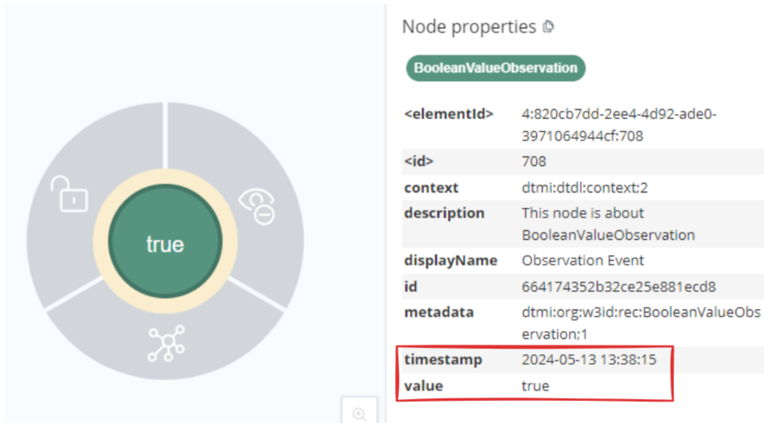

Fig. 3 Example of a boolean observation value depicting node's properties.

tial steps of a **fault detection and diagnosis** (FFD) application in data-driven SB. FDD ensures reliable operation and prevents energy waste in building systems [5]. The first step involves clustering rooms based on their time-series data or properties to establish a baseline of normal operation. Deviations from this baseline could signal potential issues, such as unexpected energy usage, malfunctions, or abnormal occupancy. Grouping rooms by occupancy characteristics is also crucial for analyzing usage patterns and predicting heating, ventilation, and air conditioning (HVAC) requirements to enhance energy efficiency [16].

1. **Identifying Occupied Rooms:** As shown in Figure 4, this query identifies occupied rooms using occupancy sensor data. In our graph-based model, a sensor node with at least one 'true' observation indicates room occupancy. We use property graphs to create a boolean attribute, 'presence,' for each room node, which is set to 'true' if any sensor has a 'true' observation and 'false' otherwise.

```
MATCH (n:Room)-[hasPoint]->(o:Occupancy_Sensor)<-[sourcePo
 ↪ int]-(m:BooleanValueObservation)
WITH n, COUNT(m) as ObsCount WHERE ObsCount > 1
SET n.presence = 1
RETURN n
```

Fig. 4 Cypher query for setting room presence based on occupancy sensor observations.

2. **Clustering Rooms Based on Occupancy:** This analysis delves into clustering building rooms based on their occupancy status on specific dates. We suggest using the GraphSAGE algorithm, which creates graph node representations as

numerical vectors by analyzing node feature information within local neighborhoods. The "presence" attribute is a feature for training the GraphSAGE algorithm in this context. Figure 5 illustrates this query in detail.

```
CALL gds.beta.graphSage.train("room_with_presence", {
    modelName:"trainModel",
    featuresProperties: ["presence"],
    aggregator:"mean", activationFunction:"sigmoid",
}) YIELD modelInfo as info
RETURN info.model as modelName
```

Fig. 5 Cypher query for training a GraphSAGE model on room presence data.

As shown in Figure 6, the next query uses the k-means clustering algorithm with k=2 to group each room node as ’empty’ or ’occupied’ based on its embedding vector. K-means is an unsupervised learning technique that aims to group data points into distinct clusters based on similarity. In this context, the algorithm identifies two clusters of nodes based on their embedding vectors, where nodes with close embedding vector representations are assigned to the same cluster.

```
CALL gds.kmeans.stream("room_with_presence" {
nodeProperty: "inMemoryEmbedding", k: 2, randomSeed: 42})
 ↪  YIELD nodeId, communityId
RETURN gds.util.asNode(nodeId).displayName AS name,
 ↪  communityId ORDER BY communityId, name ASC
```

Fig. 6 Cypher query for streaming K-means clustering results on room presence data.

The k-means clustering algorithm initially classified room nodes with an accuracy of 84%. The current k-means clustering approach may be redundant with the "presence" attribute; however, this research suggests potential for future exploration by adding other features like temperature and energy consumption. This multi-feature approach could provide valuable insights for clustering room nodes as a foundation for FFD application and real-time anomaly detection.

5 Conclusion and perspectives

This study presents a framework that implements a knowledge graph within a real use case data-driven smart building. After a state-of-the-art, the first step is identifying the most appropriate data model to gain insights depending on our needs. Then, by linking the time-series data to the standard semantics of the sensors and equipment, the resulting model facilitates the exploration of the data structure by applying graph algorithms for FFD application, serving as a preliminary decision-making tool within a digital twin framework. Future work could improve the KG by integrating additional input points and observations from dynamic data. Moreover, we could evaluate the potential incorporation of contextual data, such as BIM models. Suggestions for future work could also involve the creation of multiple subgraphs, representing "snapshots" of virtual representations of the building based on specific data selections (e.g., temperature, occupancy, energy, topology) and timeframes. GML algorithms could then find clusters, compare these subgraphs for anomaly detection, and leverage insights for FFD applications. Moreover, using a graph structure allows pathfinding algorithms to be applied to find paths between nodes. This can provide crucial insights for emergency planning and indoor navigation. Finally, integrating the DTDL and SHACL languages into the workflow has the potential to facilitate data model validation by incorporating domain constraints.

6 Acknowledgement

The authors would like to thank *Primatice* industrial partner, a subsidiary of Vinci Facilities group. *Primatice* is actively developing TwinOps[16], a platform designed to manage digital twins of real estate assets to enhance the operation and maintenance of connected buildings.

References

1. Attaran, M., Celik, B.G.: Digital twin: Benefits, use cases, challenges, and opportunities. Decision Analytics Journal **6**, 100165 (2023)
2. Baken, N.: Linked data for smart homes: Comparing rdf and labeled property graphs. LDAC2020 pp. 23–36 (2020)
3. Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y., et al.: Brick: Metadata schema for portable smart building applications. Applied energy **226**, 1273–1292 (2018)
4. Berkoben, K., El Kaed, C., Sodorff, T.: A digital buildings ontology for google's real estate. In: ISWC (Demos/Industry). pp. 392–394 (2020)
5. Blum, D., Candanedo, J., Chen, Z., Fierro, G., Gori, V., Johra, H., Madsen, H., Marszal-Pomianowska, A., O'Neill, Z., Pradhan, O., et al.: Data-driven smart buildings: State-of-the-art review. Energy in Building and Communities Programme p. 103 (2023)
6. Compton, M., Barnaghi, P., Bermudez, L., Garcia-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al.: The ssn ontology of the w3c semantic sensor network incubator group. Journal of Web Semantics **17**, 25–32 (2012)

[16] https://www.twinops.com

7. Coupry, C., Noblecourt, S., Richard, P., Baudry, D., Bigaud, D.: Bim-based digital twin and xr devices to improve maintenance procedures in smart buildings: A literature review. Applied Sciences **11**(15), 6810 (2021)
8. Deng, M., Menassa, C.C., Kamat, V.R.: From bim to digital twins: A systematic review of the evolution of intelligent building representations in the aec-fm industry. Journal of Information Technology in Construction **26** (2021)
9. Di Pierro, D., Ferilli, S., Redavid, D.: Lpg-based knowledge graphs: A survey, a proposal and current trends. Information **14**(3), 154 (2023)
10. European Union: Directive 2018/844: Amending directive 2010/31/eu on the energy performance of buildings (2018), official Journal of the European Union
11. Fierro, G., Pauwels, P.: Survey of metadata schemas for datadriven smart buildings (annex 81). Energy in Buildings and Communities Technology Collaboration Programme (2022)
12. Ghansah, F.A.: Digital twins for smart building at the facility management stage: a systematic review of enablers, applications and challenges. Smart and Sustainable Built Environment (2024)
13. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (Jun 1993). https://doi.org/10.1006/knac.1993.1008
14. Hammar, K., Wallin, E.O., Karlberg, P., Hälleberg, D.: The realestatecore ontology. In: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18. pp. 130–145. Springer (2019)
15. Lassila, O., Swick, R.R.: Resource Description Framework (RDF) Model and Syntax Specification. W3c recommendation, W3C (1999)
16. Li, T., Liu, X., Li, G., Wang, X., Ma, J., Xu, C., Mao, Q.: A systematic review and comprehensive analysis of building occupancy prediction. Renewable and Sustainable Energy Reviews **193**, 114284 (2024)
17. Lygerakis, F., Kampelis, N., Kolokotsa, D.: Knowledge graphs' ontologies and applications for energy efficiency in buildings: A review. Energies **15**(20), 7520 (2022)
18. Mateev, M.: Industry 4.0 and the digital twin for building industry. Industry 4.0 **5**(1), 29–32 (2020)
19. Poveda-Villalón, M., García-Castro, R.: Extending the saref ontology for building devices and topology. In: LDAC. pp. 16–23 (2018)
20. Qiang, Z., Hands, S., Taylor, K., Sethuvenkatraman, S., Hugo, D., Omran, P.G., Perera, M., Haller, A.: A systematic comparison and evaluation of building ontologies for deploying data-driven analytics in smart buildings. Energy and Buildings **292**, 113054 (2023)
21. Rasmussen, M.H., Lefrançois, M., Schneider, G.F., Pauwels, P.: Bot: The building topology ontology of the w3c linked building data group. Semantic Web **12**(1), 143–161 (2021)
22. Robinson, I., Webber, J., Eifrem, E.: Graph databases: new opportunities for connected data. " O'Reilly Media, Inc." (2015)
23. Rowley, J.: The wisdom hierarchy: representations of the dikw hierarchy. Journal of information science **33**(2), 163–180 (2007)
24. Sahlab, N., Kamm, S., Müller, T., Jazdi, N., Weyrich, M.: Knowledge graphs as enhancers of intelligent digital twins. In: 2021 4th IEEE international conference on industrial cyber-physical systems (ICPS). pp. 19–24. IEEE (2021)
25. Tang, S., Shelden, D.R., Eastman, C.M., Pishdad-Bozorgi, P., Gao, X.: 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 (2019)
26. Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., Sui, F.: Digital twin-driven product design, manufacturing and service with big data. The International Journal of Advanced Manufacturing Technology **94**, 3563–3576 (2018)
27. Tuhaise, V.V., Tah, J.H.M., Abanda, F.H.: Technologies for digital twin applications in construction. Automation in Construction **152**, 104931 (2023)
28. United Nations Environment Programme: Global status report for buildings and construction: towards a zero-emission, efficient and resilient buildings and construction sector (2022)
29. Wang, S., Wan, J., Li, D., Zhang, C.: Implementing smart factory of industrie 4.0: an outlook. International journal of distributed sensor networks **12**(1), 3159805 (2016)