J. L. Hernández, S. Martín, V. Marinakis and I. de Miguel, "From silos to open, federated and enriched Data Lakes for smart building data management," *2023 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*, Milano, Italy, 2023, pp. 29-33, https://doi.org/10.1109/MetroLivEnv56897.2023.10164046

From silos to open, federated and enriched Data Lakes for smart building data management

José L. Hernández
Energy division
CARTIF Technology Centre
Boecillo, Spain
0000-0002-7621-2937

Susana Martín
Energy division
CARTIF Technology Centre
Boecillo, Spain
0000-0002-1867-878X

Vangelis Marinakis
Decision Support Systems Laboratory
National Technical University of Athens
Athens, Greece
0000-0001-5488-4006

Ignacio de Miguel
Universidad de Valladolid
Valladolid, Spain
0000-0002-1084-1159

***Abstract*—Current building data is treated as silos from the different building domains. However, this provokes the lack of cross-domain data mixture to provide added-value services, mainly due to lack of interoperability. Data quality is also an issue when collecting data from buildings. The proposed data lake aims to solve these challenges by considering the whole data life-cycle to ensure minimum data quality requirements, providing high-quality services to make better-informed decisions. Heterogeneous building-related data is thus combined to enrich the information, being able to address multiple stakeholders in the smart building context. The data lake is being deployed in the DigiBUILD project, where data from 10 pilots with different purposes are collected to demonstrate the capability and benefits of its application.**

Keywords—data lake, data quality, standards, ontologies, smart buildings

I. INTRODUCTION

The majority of the European building stock was built before energy performance standards were introduced [1] and almost 97% of the EU existing buildings has low energy-efficiency performance, accounting for 40% of the total energy consumption in Europe [2]. Even when the renovation of the existing building stock seems to be the solution, aiming an energy saving of 60% or more, the reality is that the renovation rate (around 1.2% per year [3]) should rise considerably this decade to deliver on the new Fit for 55 European climate targets [4].

Within this context, not always a physical renovation of the building (e.g. improvements to envelope, heat generation, HVAC components) is feasible or assumable deriving on the need of improving the energy efficiency of the building, at its different stages, through the digitalization of the different processes. Related to this, in the scope of the EU Green Deal [4], better energy performance of buildings by increased digitalisation is earmarked as one of the policy areas together with promotion of clean energy sources. Digitalization fosters making buildings energy efficient by overarching and interconnecting energy systems and other assets to apply advance data-driven monitoring, smart assessment, prediction and optimal control strategies that should guarantee the Minimum Energy Performance Standards (MEPS) by ensuring comfort with an efficient use of resources, reducing also energy poverty.

The effect of increasing the adoption of digitalization technologies, such as Internet of Things (IoT), Artificial Intelligence (AI), Big Data or blockchain, relies on more and more data being generated nowadays as part of the building monitoring. Data concerns almost every aspect of the built environment: from how individuals and businesses use and interact with properties, to how the building's energy consumption and construction details are recorded and analysed to support informed decisions about construction and real estate processes.

On the other hand, and to accelerate and achieve the energy and environmental ambitious targets for a climate neutral building stock, it is crucial to engage a large number of stakeholders from the building life-cycle (from conceptualisation to refurbishment or demolition), and make sure that the transition towards low carbon and sustainable living is accepted and feasible for all of them.

To effectively benefit the built environment and its related stakeholders from this data-driven landscape, several technical, social and economic challenges should be overcome, starting from breaking the current silo approach thinking and continuing with the vendor lock-in relaxing and the homogenisation and standardization of the data acquisition and storage through the application of technical and semantic interoperability enablers. Only integrating and enriching data from multiple data sources, and considering the needs and interest of the stakeholders involved in the building domain, effective and holistic decision-making can be achieved.

With this aim in mind, this paper proposes and presents an open and enriched Data Lake able to combine dynamic, static and contextual data from multiple and heterogeneous data sources to provide more valuable information through Data Marts according to different building domains and based on Business Intelligent technologies. This Data Lake also applies federated conceptualisation through the application of automated model and data quality checking to ensure high quality data is exposed to upstream analytics and computing tools using uniform and well-defined interfaces. All in all, the Data Lake supports the entire data life-cycle management processes, from data ingestion and pre-processing to quality checking and data normalization, to expose multiple data views to analytics, services or any other tool built on top of it.

This Data Lake is being implemented under the umbrella of the DigiBUILD project [5], whose main aim is to provide high-quality services for the digital transformation of buildings, according to the EU Green Deal specifications [4].

The final result is an open, interoperable and cloud-based toolbox based on a federated and enriched Data Lake.

II. BACKGROUND AND PROGRESS BEYOND THE STATE OF THE ART (SOTA)

Current data management strategies in the building stock are based on traditional silo approaches (Fig. 1), where stakeholders manage their own data. Moreover, existing buildings present low level of digitalisation, heterogeneous data sources and low data quality, while the occupants are not included in the lifecycle.

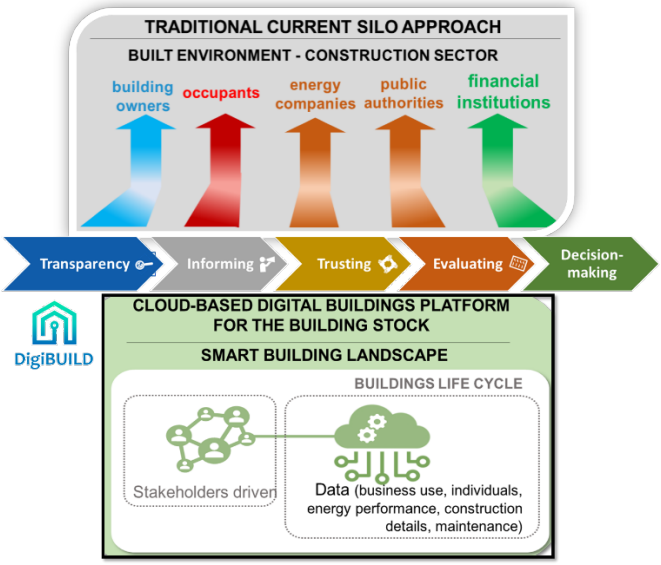

Fig. 1. Tradicional building data silos vs DigiBUILD data lake

These facts can be translated into several challenges [5]:

- Multiple stakeholders are participants within the built environment, with different data needs because of the use of these data.
- Traditionally, buildings are isolated elements without collaboration among them, which is directly reflected in the data collection approaches, providing multiple data repositories, non-harmonised and lacking of interoperability.
- Data protection and security is an open issue that ends in unreliable quality and limited accessibility to the energy-related data.

Many efforts have been put in place to overcome the aforementioned challenges, however, still, the heterogeneous nature of the building related data makes the integration very challenging [6]. Multiple domains require the proper data treatment methodology, e.g. BIM (Building Information Modelling) or dynamic data timeseries. Moreover, there is a clear fragmentation in the AEC (Architectural, Engineering and Construction) industry, as stated in the first challenge. The main result is the requirement to interface multiple communication protocols, establish ad-hoc data formats and ontologies with diverse targets [7]. It is true that efforts are made to homogenise the communication standards, but IoT protocols are still vendor dependent [8].

Day-by-day, buildings generate more and more data, which need to be processed by data management platforms [9], but the quality of the data remains the same [10]. There is a clear gap in the reliability of the services due to the low quality [11]; therefore, the ability to process and trace data errors is a key topic [12].

BIM can establish the grounds for a harmonisation of the data samples to allow the interaction of the physical environment with data management systems [13]. Better-informed decision-support systems should benefit from the building characteristics to represent building assets [14].

A. *Progress beyond the SotA*

A step forward is clearly necessary for digitalisation of smart buildings, merging heterogeneous data, increasing data quality and putting the stakeholders in the centre of the building. In this sense, this paper presents an open, federated and enriched data lake to solve many of the previous challenges in the current practices.

First of all, the data lake provides a dynamic and adaptable interoperable framework based on the existing standards according to the energy-related application [7]. Some examples are SAREF, BOT, IFC or BRICK. While the current SotA selects a single ontology, the proposed data lake merges data-sets by domains.

Secondly, three interoperability levels are proposed [9]:

- Southbound, where heterogeneous data samples are interfaced by middleware and/or data broker approaches to homogenise data before being stored in persistent databases.
- Northbound to set the ways to share data among the stakeholders and, then, ensure a user-centric approach.
- Semantic, based on the previous dynamic and adaptable ontologies.

Thirdly, the data lake concept considers the whole data life cycle (see Fig. 2), where quality methodologies are applied. Hence, data protection and security are considered since gathering processes. Additionally, not only the data collection is conducted, but also the minimisation of the error propagation is focused.

III. METHODOLOGY: DATA LIFE CYCLE

As anticipated, the Data Lake covers the full data life cycle, which is depicted in Fig. 2 [9]. From the data acquisition to the exploitation, the transformation and treatment mechanisms are part of the methodology applied within the Data Lake to assure interoperability. This methodology is driven by the three interoperability levels that were described above. In this way:

- Southbound is composed by the data acquisition and data ingestion procedures, implemented through the drivers and APIs in charge of interfacing the field level. Raw data is thus gathered in the specific protocols, which is orchestrated to provide synchronisation in the data sampling. Both static and dynamic data are considered in this stage, compiling, on one hand, timeseries from sensors about the building operation (e.g. indoor temperature, energy used for heating…) and, on the other hand, contextual data from the building (e.g. Building Information Model – BIM).
- Semantic is the second stage of the methodology, composed by multiple steps. Firstly, the integration of

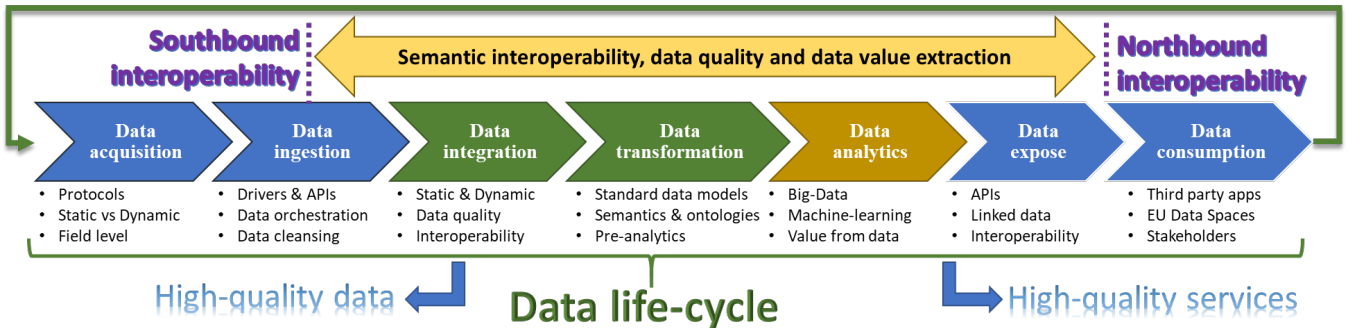

Fig. 2. Data life cycle

data coming from the various domains (sensors, BIM, users or digital logbooks, among others) that performs the data quality checks in several dimensions (completeness, reliability, consistency, accessibility, accuracy, relevance, timeless), increasing the quality and, hence, the credibility of the high-level services, making better-informed decisions. Additionally, this is the first level within the semantic interoperability by merging data samples.

Next phase is the data transformation, where semantics and ontologies are applied to transform the merged data sets into standard data models to represent data. Here, a dynamic and adaptable interoperable framework is established to accommodate data according to the building objectives. That is to say, depending on the Smart Building services, data models are shaped to fit into the building characteristics (i.e. district heating services do not publish similar data than comfort parameters).

Finally, a set of analytics are applied to obtain value from raw data. Big-Data techniques are then applied to take advantage of the data model benefits, which is named linked data (see section IV-A). The application of these techniques allow extracting, for instance, the relation between data samples and building assets, such as geolocation of heat pumps and how the use of such a heat pump could affect the temperature of an adjacent space or thermal zone of the building. Moreover, learning about the behaviour of the energy systems, such as energy inertia to heat/cool a space or forecast the renewable production of photovoltaics. These are the pre-analytics in form of data marts, explained in the next chapter.

- Northbound, which is the last stage of the methodology and focuses on the data exposing and sharing through intelligent querying mechanisms. These are APIs with the ability of merging and filtering data from multiple data marts, according to the user criteria.

The main benefits of applying this methodology can be classified in two main aspects:

- Reliability and Credibility by applying the data quality methodology in the data life cycle. This reduces and corrects data errors, not only in the data gathering process, but also in the propagation. Thanks to the multi-dimension criteria, data quality are improved in terms of: (1) Data gaps, reduction of missed data; (2) Outliers that are produced by values that differ from the expected measurements, thus, reducing the uncertainties of data analytics; (3) Consistency and accuracy by removing duplicates, aligning multi-source measurements (e.g. date/time in timeseries data from various sources) and cleaning ambiguous values. (4) Model check, which is focused on static data (i.e. BIM). Errors in the building modelling are usual and propagated to the tools for the building life-cycle management. These mistakes should be prior detected and solved whenever possible.
- Interoperability: Better-informed decisions mean using combined data from heterogeneous sources (e.g. sensors, energy performance certificates databases, digital logbooks, BIM, CityGML, Level(s), etc.), but the lack of interoperability is an issue to merge these datasets [7]. The data life cycle methodology makes sure interoperability is covered at multiple levels, including the dynamic and adaptative data models to accommodate data into the requirements of the use cases or services to be deployed.

IV. Data Lake Definition

Driven by the motivation of moving from traditional silo approaches to digital and high-quality data-driven Smart Buildings, an open, enriched and federated Data Lake is being developed in the context of the DigiBUILD project [5].

A Data Lake is considered as a (not necessarily centralized) repository where structured, semi-structured and unstructured data can be stored at any scale [15]. In this sense, and differentiating the Data Lake from a traditional Data Warehouse, different types of analytics can make use of the data available in the Data Lake without the need of a previous structuration or filtering of this data. To facilitate data searching, the Data Lake counts on a Data Catalogue Service able to catalogue and index the data, and the proper connectors to expose the data to analytics and machine learning tools in upper layers.

The Data Lake is conceived to offer a holistic people-centric data framework for gathering and managing building data, guaranteeing minimal syntactic (communication protocols) and semantic (common data representation) interoperability.

At a first glance, the Data Lake is fed by (Fig. 3):

1. Open, real-time (i.e. dynamic) and reliable building data from multiple sources, such as (smart) equipment connected to the building.
2. Contextual and static data related to non-energy assets of the building, such as information related to geometry of the buildings (IFC, CityGML, cloud points, etc.), user

behaviour during building operation, Digital Logbooks, energy performance certificates (EPCs), or Level(s).

3. Other external data sources (weather forecast, climate, market data).

Once analytics and third-party services are run over the Data Lake, the results will be also stored in the Data Lake to enrich the information existing about the buildings and to re-calibrate the models to increase their accuracy.

Gathering this data, the Data Lake is conceived to manage data in a proper and adaptable way to drive more robust, improved and consistent monitoring of building stock energy performance, and through the whole value chain.

An added value of the Data Lake, trying to avoid the mobilization of big amounts of data, is the creation of a set of Data Marts, containing the required dynamic, static and contextual linked data per building domain, such as thermal energy, electricity, electrical vehicle, comfort, EPCs or Smart Readiness Indicator (SRI). These Data Marts expose linked, enriched, pre-processed and high-quality data to the AI-based analytics, Digital Building Twins and other initiatives and open data spaces by using Business Intelligence (BI) and intelligent querying.

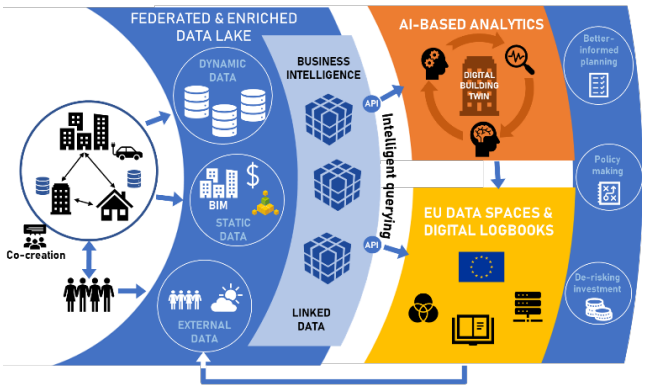

Fig. 3. Conceptual schema of the Data Lake and high-level analytics

Under this vision, the DigiBUILD Data Lake is considered "Open" because it makes data available through uniform interfaces to upstream analytics and computing tools, linking it to other existing Open Data Spaces. In addition, the Data Lake is consider "federated" because it supports automated model and data quality checking services, improving different data quality dimensions in terms of completeness, consistency or accuracy. Finally, the Data Lake is considered "enriched" because of the integration of data and metadata from multiple and heterogeneous data sources, supported by Building BI applied to different Data Marts containing linked data per building domain.

A. Linked data

To make high-quality data available, the Data Lake will combine dynamic, static and contextual data to provide more valuable information through the Data Marts per building domain. The dynamic database (as part of the Data Lake) will exploit timeseries and relational databases concepts to represent dynamic data samples with contextual information. Data Warehouse concepts will be also considered by using fact tables as timeseries and dimensional tables as relationships. Part of these dimensional tables will create the link with static data (Fig. 4), such as BIM, in order to provide advance analytics, as for example, being able to determine the Indoor Environmental Quality (IEQ) comfort parameters per space of a building by georeferencing sensors based on the information available in the BIM model.

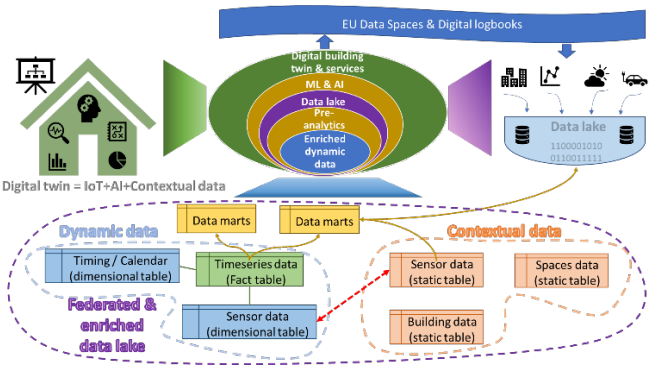

Fig. 4. Linked data for the creation of the federate and enriched Data Lake

Dynamic and static data will co-exist, enriched with ontologies and semantics that provide metadata to the data, to be exploited for high-quality energy services. These ontologies (such as SAREF, Brick or Haystack) will support data interoperability, creating open, linked data to be shared through intelligent querying services, providing better availability of big-data analysis.

Merging and combining these data sources, the Data Lake promotes the capability to obtain valuable information from data through the Data Marts and pre-analytics, fostering the creation of high-quality energy and non-energy services.

A main advantage of this two-stage approach for data integration and availability is the reduction of data load in the communication networks. This is achieved by just sharing Building BI data and not the full data stock, as well as limiting the computational load of the AI-Digital Twin calculations, by providing pre-calculated analytics, such as aggregated consumptions and comfort values. In this way, the enriched dynamic repositories will only integrate the necessary static data, for example by extracting specific model views, following the MVD-Model View Definition paradigm [16], and dynamic data, relying on the local databases from the pilots, to produce upward federation.

By merging Data Marts from different domains, the federated Data Lake will provide a holistic view of the building representation, performance and/or operation form cross-cutting domains (thermal energy, electricity, EV, comfort, performance, EPCs, SRI). The use of federated-data completeness checking, e.g. using the Shapes Constraint Language (SHACL) for static data, will ensure all required data is present, prior to making these available to upstream applications.

B. Benefits for Smart Buildings

Although several researches have been conducted [17][18], nowadays, data is exponentially growing, including new sources and novel ontologies. Moreover, Artificial Intelligence and machine-learning techniques have found new paradigms in the building sector, which is continuously changing. Having said that, data mining, treatment and management methods should adapt new trends. Indeed, within the research made in [17], only IFC was considered as data model, while [18] extends the analysis including sensor data, but still focused on BIM availability.

The Data Lake described in this paper benefits the integration of multiple and heterogeneous data sources in

multiple domains, thanks to the Building Business Intelligence and the use of data marts. The use of domain-driven analytics, on the one hand, provides users with relevant data to make decisions and, on the other hand, exploit the correlation in the data-sets.

Additionally, one of the major benefits of the use of this approach lies in the data quality. Previous researches do not consider data quality issues, then, driving to decisions taken based on non-complete or low accurate data. Reduction of data errors increment the credibility of the decision-making process with better-informed users.

V. Conclusions and Future Work

Data is considered the oil of the XXI century, but its availability and quality are limited. Sometimes, even though data is accessible, it requires enrichment from additional sources, which is not a trivial task and requires treatment mechanisms. For that end, the Data Lake presented in this paper aims the data gathering of multiple and heterogeneous data sources that can provide a holistic perspective of the building operation. The main objective is to provide high-quality data to the users in order to make better-informed decisions.

To do so, a data life cycle methodology has been defined, with the goal of assuring the reliability, credibility, interoperability, privacy and security of data. By assuring these aspects, buildings' stakeholders can rely on the data analytics and data-driven and performed better-informed decisions about building management strategies (e.g. de-risking financial projects in terms of energy efficiency). The Data Lake approach also allows the creation of data-driven business models, providing more accurate and trustworthy value (e.g. ESCO models where energy prices are based on performance calculation and energy savings could be calculated based on wrong data).

With respect to the future lines, first of all, it should be highlighted the project is in the very early step, defining requirements; therefore, its implementation and validation in key in the process. Additionally, external sources combination is another research line to increase the availability of data sets, such as dynamic energy prices to correlate energy uses with dynamic pricing to calculate analytics in this sense. Finally, buildings complexity should be accounted because the configuration of the heating and cooling systems, which are not always represented in static models, being necessary its identification with different types of data models to extend the current practices.

Acknowledgements

The authors of this paper would like to thank the European Commission for funding the project DigiBUILD under the GA#101069658. Additionally, they would like to also thank the DigiBUILD consortium for the work and contributions that support the creation of this publication.

References

[1] European Commission, https://ec.europa.eu/energy/eu-buildings-factsheets-topics-tree/building-stock-characteristics_en, accessed on 21st February 2023.

[2] Building stock observatory, https://ec.europa.eu/energy/news/building-stock-observatory-new-database-european-building-stock-and-its-energy-performance_en, accessed on 21st February 2023.

[3] European Commission, Evaluation of the Energy Performance of Buildings Directive 2010/31/EU, 2015.

[4] European Green Deal, https://www.consilium.europa.eu/en/policies/green-deal/fit-for-55-the-eu-plan-for-a-green-transition/, accessed on 21st February 2023.

[5] DigiBUILD project. 2022. https://digibuild-project.eu/, accessed on 20th February 2023. GA#101069658, doi: 10.3030/101069658.

[6] J. Koh, S. Ray, and J. Hodges, 'Information Mediator for Demand Response in Electrical Grids and Buildings', in 2017 IEEE 11th International Conference on Semantic Computing (ICSC), Jan. 2017, pp. 73–76. doi: 10.1109/ICSC.2017.26.

[7] F. De Andrade Pereira et al., 'Towards semantic interoperability for demand-side management: a review of BIM and BAS ontologies', presented at the EC3 Conference 2022, 2022, vol. 3, pp. 0–0. doi: 10.35490/EC3.2022.154.

[8] F. de Andrade Pereira, C. Shaw, S. Martín-Toral, R. Sanz Jimeno, D. Finn, and J. O'Donnell, 'Exchange requirements to support demand side management using BIM and Building Automation System domains', presented at the 2021 European Conference on Computing in Construction, Jul. 2021, vol. 1, pp. 0–0. doi: 10.35490/EC3.2021.202.

[9] J. L. Hernández, R. García, J. Schonowski, D. Atlan, G. Chanson, and T. Ruohomäki, 'Interoperable Open Specifications Framework for the Implementation of Standardized Urban Platforms', Sensors, vol. 20, no. 8, Art. no. 8, Jan. 2020, doi: 10.3390/s20082402.

[10] C. Duvier, D. Neagu, C. Oltean-Dumbrava, and D. Dickens, 'Data quality challenges in the UK social housing sector', International Journal of Information Management, vol. 38, no. 1, pp. 196–200, Feb. 2018, doi: 10.1016/j.ijinfomgt.2017.09.008.

[11] S. Y. Teng, M. Touš, W. D. Leong, B. S. How, H. L. Lam, and V. Máša, 'Recent advances on industrial data-driven energy savings: Digital twins and infrastructures', Renewable and Sustainable Energy Reviews, vol. 135, p. 110208, Jan. 2021, doi: 10.1016/j.rser.2020.110208.

[12] N. Hossein Motlagh, M. Mohammadrezaei, J. Hunt, and B. Zakeri, 'Internet of Things (IoT) and the Energy Sector', *Energies*, vol. 13, no. 2, Art. no. 2, Jan. 2020, doi: 10.3390/en13020494.

[13] N. Luo, M. Pritoni, and T. Hong, 'An overview of data tools for representing and managing building information and performance data', Renew. Sust. Energ. Rev., vol. 147, p. 111224, Sep. 2021, doi: 10.1016/j.rser.2021.111224.

[14] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, 'A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data', Front. Energy Res., vol. 9, p. 652801, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[15] Data Lakes: The Definitive Guide. Paul Singman, March 22, 2021. Available in https://lakefs.io/blog/data-lakes/, accessed on 24th April 2023.

[16] Model View Definition (MVD) – An Introduction. Data Standards, Industry Foundation Classes (IFC). https://technical.buildingsmart.org/standards/ifc/mvd/, accessed on 24th April 2023.

[17] Mo, K., Menzel, K., & Hoerster, S. (2013). Development of an IFC-compatible Data Warehouse for Building Performance Analysis.

[18] Menzel, K., Törmä, S., Markku, K., Tsatsakis, K., Hryshchenko, A., Lucky, M.N. (2022). Linked Data and Ontologies for Semantic Interoperability. In: Daniotti, B., Lupica Spagnolo, S., Pavan, A., Bolognesi, C.M. (eds) Innovative Tools and Methods Using BIM for an Efficient Renovation in Buildings. SpringerBriefs in Applied Sciences and Technology(). Springer, Cham. https://doi.org/10.1007/978-3-031-04670-4_2.