Using Extraction, Transformation and Loading procedures for digitalisation of buildings

José L. Hernández
Energy Division
CARTIF technology centre
Boecillo, Spain
josher@cartif.es
0000-0002-7621-2937

David Arévalo
Energy Division
CARTIF technology centre
Boecillo, Spain
davare@cartif.es
0009-0009-8683-6741

Susana Martín
Energy Division
CARTIF technology centre
Boecillo, Spain
susmar@cartif.es
0000-0002-1867-878X

Kyriakos Katsigarakis
IEDE
University College London
London, UK
k.katsigarakis@ucl.ac.uk
0000-0002-2748-4506

Georgios N. Lilis
IEDE
University College London
London, UK
g.lilis@ucl.ac.uk
0000-0002-0642-5291

Dimitrios Rovas
IEDE
University College London
London, UK
d.rovas@ucl.ac.uk
0000-0002-5639-6783

Ignacio de Miguel
Universidad de Valladolid
Valladolid, Spain
ignacio.miguel@tel.uva.es
0000-0002-1084-1159

***Abstract*—The digitalisation of the building stock necessitates the integration of a wide array of digital and non-digital data sources into a cohesive framework that adheres to standardized data formats. Achieving this integration involves employing various extraction, transformation, and loading processes. These processes play a crucial role in converting raw data collected from building sites into instances that align with the specified unified format. This work delves into extraction, transformation, and loading methods utilized across nine pilot building sites situated in different countries, each marked by substantial data diversity. The heterogeneity among data sources and, consequently, datasets, is effectively addressed by a customized gathering process. This process incorporates static data to enhance the overall quality, enabling better-informed decision-making. The result is a harmonized building data repository with 10 use cases and more than 8000 data points, facilitating the application of intelligent services for energy-efficient management strategies. Enrichment of data is also achieved by synchronization approaches to ensure the coherence of the data.**

***Index Terms*—Data lake, Extraction, Transformation and Loading (ETL), interoperability, data syncing, timeseries, static data.**

I. Introduction

Digitalisation is of paramount importance in today's rapidly evolving landscape, fundamentally transforming the way of data collection, processing, and sharing. One key facet of this transformation lies in the development and widespread adoption of big data. Big data, with its vast and intricate datasets, empowers organizations to derive valuable insights, make better-informed decisions, and innovate across various sectors [1]. The ability to capture and analyse massive amounts of data in real time not only enhances operational efficiency, but also unveils patterns and trends that might otherwise go unnoticed. From optimizing business processes to advancing scientific research, big data fuels innovation fosters strategic planning and ultimately propels progress. The synergy between technological advancements and the utilization of big data stands as a catalyst for efficiency, innovation, and informed decision-making in the interconnected world [2].

The rise of this digital transformation is fueled by the growing adoption of cutting-edge information and communication technologies (ICTs), such as the Internet of Things (IoT) and/or artificial intelligence (AI). Data permeates virtually every facet of the built environment, encompassing how individuals and businesses utilise and engage with properties [3]. It extends to the recording and analysis of a building's energy consumption and construction details, facilitating better-informed decisions in construction and real estate processes. Harnessing data for decision-making and embracing digital enhancements can significantly enhance operational efficiencies at a minimal cost.

Ensuring transparency and trust in the decision-making process to achieve the objectives of the Energy Performance of Buildings Directive (EPBD) [4] requires a meticulous validation and detection of gaps, incorrect, or inaccurate data across the entire value chain of building monitoring. This crucial step is integral to the directive's goals of promoting low energy usage, minimizing carbon footprint, optimizing thermal comfort, and evaluating air quality.

To effectively leverage the potential of this data-driven landscape for the built environment and its stakeholders, various technical, social, and economic challenges must be addressed. The main challenge involves overcoming the prevailing silo approach, which leads to vendor lock-in when integrating data from diverse and heterogeneous sources [5]. Digital solutions play a pivotal role in achieving the sustainability

This research has been funded by the Horizon Europe programme, GA #101069658.

targets by delivering energy savings ranging between 10-15%, through the implementation of optimal control strategies and the indirect savings resulting from users being better informed [6]. Furthermore, there is a need for standardization throughout the data lifecycle, encompassing acquisition, sharing, and storage. This addresses interoperability issues and facilitates the seamless merging of data from multiple building domains, thereby enhancing data enrichment [7].

Addressing these challenges is crucial for unlocking the full potential of the data-driven landscape in the built environment. This work then advances with respect to the current practices to harmonise dynamic datasets and synchronise with static metadata in order to provide a unified data and store information in common data models. This procedures then fosters the engagement of stakeholders involved in different aspects of the co-created interoperable services and tools that make use of the uniform data space yielded by the federation of the data to be able to make better performance monitoring, have a more trustworthy decision-making for assessment and planning of building infrastructure, policy making and re-risking investments, in the end, an energy efficiency-focused set of services and tools.

This work is executed under the European project DigiBUILD [8], which is focused on creating a uniform data space fed by an abundant number of data sources using different protocols to access them. By integrating state of the art big data techniques and methods, data collection mechanisms are implemented. DigiBUILD project counts on nine pilots in different locations (U.K., Greece, Spain, Italy and The Netherlands) to validate the developments across multiple climates and building topologies in Europe.

The rest of the paper is organised as follows. Section II provides the concept of the data lake that supports the digitalisation process. Section III describes the Extract, Transform and Loading (ETL) processes to gather heterogeneous data. Section IV includes the integration of metadata to enrich the dataset obtained from the data sources. Finally, Section V summarises the insights and future work.

II. Methodology of digitalisation

Extract, Transform and Loading (ETL) procedures are a crucial step in data integration that involves extracting data from various sources, transforming it into a consistent and usable format, and loading it into a target database or data warehouse [9]. ETLs play a pivotal role in facilitating data-driven decision-making by ensuring that information is accurately collected, processed, and made available for analysis. Innovations in ETL processes have been driven by advancements in cloud computing, big data technologies, and the growing complexity of data sources. Cloud-based ETL solutions offer scalability, flexibility, and cost-effectiveness [10], allowing organizations to adapt to changing data needs [11]. Additionally, modern ETL tools leverage machine learning and automation to enhance data cleansing, transformation, and enrichment, reducing manual efforts and improving efficiency [12]. Real-time ETL processes have become more prevalent, enabling organizations to work with up-to-the-minute data for timely insights. The evolving landscape of ETL reflects a commitment to addressing the challenges posed by the ever-expanding volume and diversity of data in today's digital age.

Within this work, these ETLs are part of a data lake architecture that represents a significant stride in addressing prevailing challenges. Firstly, it places a clear emphasis on the data gathering process, with a central focus on robust data quality methodologies. Secondly, it advocates for a dynamic and adaptable interoperable framework grounded in existing standards specific to energy-related applications. Thirdly, the perspective on interoperability is distinctly outlined across three levels [2]:

- Southbound: At the field level, data sources interface with consideration for current protocols and data formats, establishing data brokers and synchronization mechanisms to homogenize data before persistent storage.
- Northbound: Interfaces for data sharing are designed to provide stakeholders and intelligent services with information based on Business Intelligence approaches.
- Semantic: Combination of dynamic and static building datasets through adaptable data models tailored to building requirements and the services to be deployed. This ensures a comprehensive and flexible integration of building data for enhanced functionality.

This approach is depicted in Fig. 1, where the data lake concept is illustrated, so that digitalisation process could take place. The first step right after extracting the data from the many different sources is unifying the format that data comes in. This is carried out using tailored ETLs developed in the Pentaho Data Integration tool.

The ETLs serve as the key drivers facilitating data gathering from both dynamic (i.e., timeseries) and static (i.e., contextual) repositories in the pilots. In Fig. 2, the integration of data sources into the digitalization schema is illustrated. Deployed alongside are dedicated data storage repositories, necessitating synchronization mechanisms to maintain data coherence. Atop these repositories, data marts utilize business intelligence (refer to Fig. 1) to aggregate and amalgamate data from

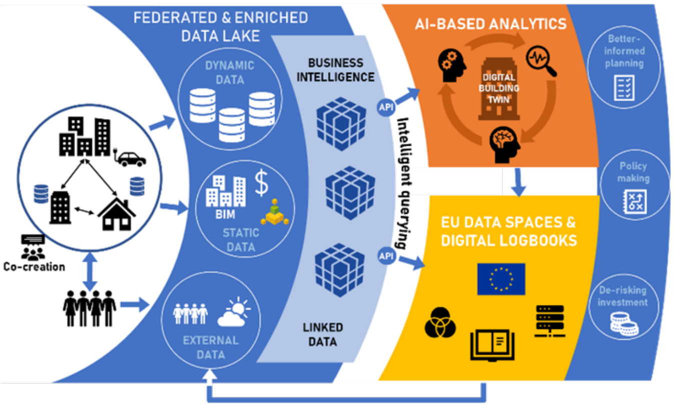

Fig. 1. Conceptual schema of the digitalisation process

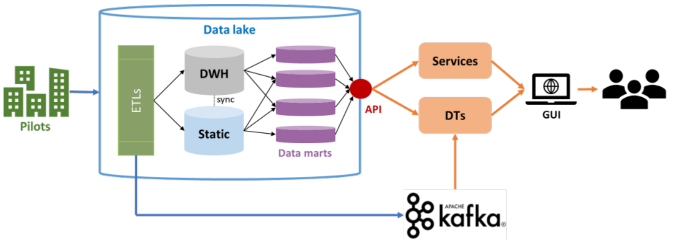

Fig. 2. ETLs within the building digitalisation process

these sources. Consequently, these data marts offer intelligent querying mechanisms, enabling the efficient sharing of data.

III. ETLs for Dynamic Data Collection in the Pilots

Data sources are diverse and heterogeneous from pilot to pilot (i.e., test cases where the ETLs have been deployed). Table I summarises the multiple datasets provided in each one of the pilots, with the communication protocol used in the data extraction, as well as the data format. This heterogeneity stands out as a primary challenge and constraint during the data gathering process. It demands a customized implementation of interfaces to access raw data.

This approach is visually represented in Fig. 3, depicting three distinct data pipelines. The blue pipeline pertains to dynamic data, involving interfacing with the pilot repository, filtering data, and extracting measurements for storage in the common data model. The orange pipeline begins with the filtered dataset, extracting metadata to synchronize with static repositories. The green stream shares real-time data, enabling services to dynamically update conditions. It should be noted the capability of managing real-time data processing. According to the interfaces in TABLE I, velocity is one of the key aspects. MQTT (Message Queuing Telemetry Transport) provides streaming data, while FTP (File Transfer Protocol) is accessible once per day. The use of the methodology proposed in this manuscript allows multiple processing timing, enhancing the capability of data analysis in real-time.

It is not just about establishing access points; the format also needs harmonization to achieve a standardized representation of information. In this context, each pilot confronts a unique set of challenges, as elaborated in the following sections.

A. UCL pilot (United Kindom)

UCL manages the ingestion of data sources from an MQTT broker, where an approximate volume of 3800 data-points is published in plain text format. The primary objective of this pilot is the effective management of the substantial data influx arising from this multitude of data-points. To accomplish this, the ETL interfaces the MQTT broker to consume streaming data. Presently, the configuration capabilities of the MQTT Consumer impose constraints, compelling to adopt a specific approach. This entails subscribing to all topics and subsequently employing topic filtering to selectively extract the relevant data. Following this step, the separation of metadata and timeseries data takes place.

TABLE I
DigiBUILD pilot interfaces to gather datasets

Pilot	**Interface details**		
	Dataset	***Protocol***	***Format***
UCL (U.K.)	BMS (Building Monitoring System)	MQTT	Text
	EMS (Energy Management System)		
	Light Controls		
	Access Controls		
	Occupancy		
EDF (France)	Ethera	Nemocloud API	JSON
	Ellona	Ellonasoft API	
	Wattsense	Wattsense API	
IASI (Romania)	3PhaseMeters	InfluxDB	CSV
VEOLIA (Spain)	Building EMS	FTP	CSV
	District EMS		
EMOT (Italy)	Charging Stations	API	JSON
	PV and building		
	eV data		
FOCCHI (Italy)	PV data	SolarEdge API	JSON
	Comfort and energy data	JotMotiqa API	
	Comfort and energy data	MQTT	Text
	Occupancy	Google Calendar API	CSV
	Energy Consumption	Google Drive API	
HERON (Greece)	Building data	API	JSON
FVH (Finland)	BEMS	FTP	CSV
IEECP (The Neth.)	Building data	MQTT	Text
		Netatmo API	JSON
NTUA (Greece)	BMS	PostgreSQL	Text

B. EDF pilot (France)

There are three distinct data pipelines, each corresponding to a different interface with unique authentication and

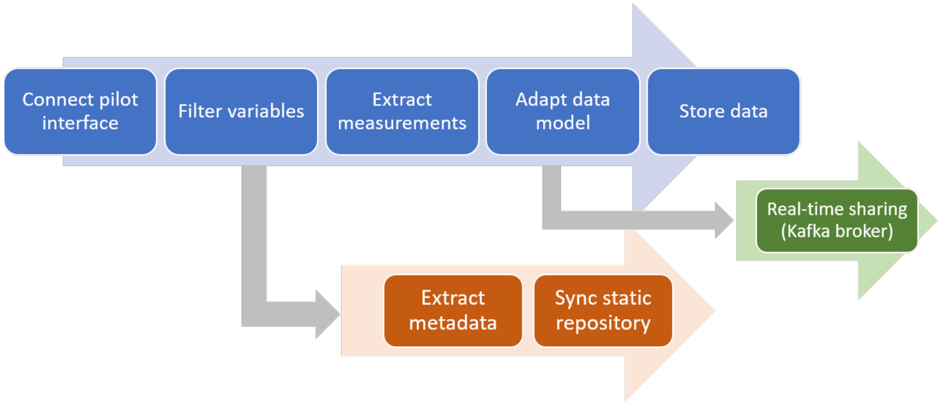

Fig. 3. ETL approach for the data collection from pilots

identification methods. For the Nemocloud API, credentials are utilized, which are provided earlier in the HTTP request body and the authorization header populating it from the ”digest” process, acquiring a session token. Subsequently, it is possible to navigate through available devices and variables before making the data request. In the case of the Ellonasoft API, an authentication token is obtained by including the credentials in the HTTP request body, and data extraction occurs in a following HTTP request. As for the Wattsense API, authentication is achieved through the authentication tab when requesting data. The data, extracted in JSON format from each interface, undergoes separation into metadata and timeseries data in each pipeline and is ultimately stored in the data warehouse.

C. IASI & SITTA pilot (Romania)

The ETL developed for the IASI & SITTA pilot introduces a different approach to data extraction. Prior to the transformation steps, authentication and data extraction are executed through a Python script within the job containing the ETL. This necessity primarily stems from the absence of the InfluxDB version 3 plugin in Pentaho’s marketplace. In this script, it is instantiated an InfluxDB client with the provided credentials and TLS certification. Subsequently, SQL queries are generated and executed to retrieve electricity consumption and air quality data, storing the results in a CSV file. Finally, within the ETL process, metadata and timeseries data are separated and stored in the Data Warehouse.

D. VEOLIA pilot (Spain)

The VEOLIA pilot comprises two distinct complexes. Consequently, two separate ETLs have been developed, despite the fact that the applied transformations are identical for both sites. In this scenario, a unique challenge arises when establishing the connection for the extraction interface. The extraction process is carried out through SFTP. VEOLIA daily publishes a CSV file on their FTP server every day, encompassing over 750 variables for on pilot and 15 for the second one. During the transformation phase, the focus is on filtering and formatting only the relevant variables, which are then separated into the metadata and timeseries data and stored in the Data Warehouse.

E. EMOT pilot (Italy)

In the EMOT pilot, access to a public API is granted, enabling to request data from three distinct pipelines: building energy consumption, photovoltaic production, and electrical vehicle charging stations. The data extraction is facilitated through a straightforward HTTP request, lacking the capability to specify the desired time period for data retrieval. This temporal filtering is assumed and executed by the ETL process. Finally, the extracted data from all three pipelines is stored in the Data Warehouse.

F. FOCCHI pilot (Italy)

In addressing the challenges posed by the FOCCHI pilot, it is grappled with the intricacies of managing six distinct interfaces for data extraction, each adhering to its format. Pertaining to photovoltaic production, power, and energy data, an API delivering JSON-formatted data is accessed, which was subsequently merged into a unified flow. Additionally, this pilot encompasses data from one of FOCCHI’s factory complex offices, retrievable monthly through the Google Drive API in CSV format. For the ground floor room, data is ingested via an MQTT broker in plain text, focusing on comfort and lighting parameters. In contrast, the first-floor room employs an API-based interface for retrieving comfort, lighting, and energy consumption parameters in JSON format. Furthermore, a schedule from a meeting room in the same office building serves as a metric of occupancy through the Google Calendar API in JSON format.

G. HERON pilot (Greece)

The connection process in the HERON pilot, illustrated in Fig. 4, followed a familiar path: authentication was achieved through credentials embedded in the body of an HTTP request, resulting in the acquisition of a session token. The particular challenge in this pilot lay in the formatting of timeseries data. The sensor devices comprise smart meters and smart plugs, measuring three phases for power, energy, returned energy, and

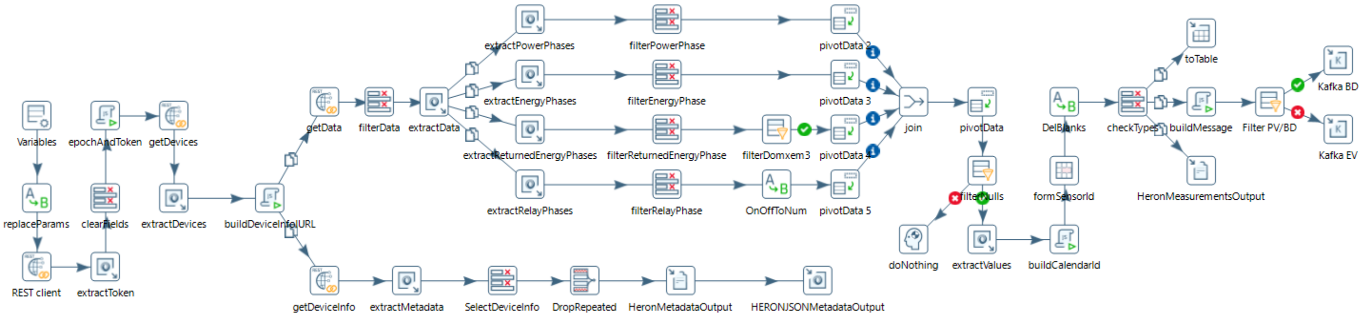

Fig. 4. ETL for the HERON pilot where the extraction of data requires additional filtering processes

relay. Some devices lack data for returned energy and relay, and the nulls they generate are appropriately filtered out. Metadata retrieval simply involved requesting device information. In the end, after merging and pivoting the various measured magnitudes, the data is stored in the Data Warehouse.

H. FVH pilot (Finland)

In the context of FVH, a parallel scenario to VEOLIA emerges concerning data extraction, conducted daily through SFTP. The outcome involves obtaining a compressed CSV file for timeseries data and a decompressed CSV file for metadata. Both the file transfer and decompression operations are executed within the job file. The subsequent transformation steps mirror those employed in previous cases, ending in the storage of data within the warehouse. Following this process, the decompressed CSV file is deleted, and the compressed version is relocated to a distinct directory.

I. IEECP pilot (The Netherlands)

For this particular pilot, two distinct pipelines are interface, employing MQTT and an API connection, respectively. In the MQTT consumer, the subscriptions extend to electrical meter and relay topics, with messages lacking a specific timestamp. Assuming immediate production and minimal consumption delay, the timestamp is appended using the system's time reference. In the second pipeline, an external list of devices and their credentials, encapsulated in a JSON file, is provided. For each device, a refreshed token is requested to facilitate authentication though an API request. Subsequently, timeseries data is requested and appropriately formatted, while metadata extraction relies on information sourced from the stations. The outputs from both pipelines find their storage destination in the data warehouse.

J. NTUA pilot (Greece)

In the NTUA pilot, data extraction involves reforming SQL requests directed at a PostgreSQL database. The primary challenge in this scenario lies in effectively managing the substantial volume of these requests and manually specifying the time period for each. To address this, two distinct inputs are devised. The first one assimilates data from over 20 diverse tables, encompassing information on air conditioning and lighting. Concurrently, the second flow draws data from six distinct tables, focusing on comfort parameters. Following the extraction of metadata from both pipelines, timeseries data is seamlessly merged and processed before finding its storage destination in the data warehouse.

IV. Integration of Contextual Data

Semantic web technologies are utilized to connect dynamic data with static contextual data. At each pilot site, a semantic graph is created. This graph consists of semantic graph nodes, each corresponding to physical sensing devices or virtual simulation points in a one-to-one mapping relationship. These semantic graph nodes adhere to the core module of the ontological scheme known as the DigiBUILD ontology, which has been documented online: [13] and uses classes from real estate core (`rc:`) and BRICK ontologies (`brick:`). The core module of this ontology is displayed in Fig. 5. In this diagram, the connection between the references to static data (`brick:point`) and the references to dynamic data entries (`:Timeseries`) is emphasized with a red dashed rectangle in the bottom right section. Essentially, the `:Timeseries` class contains information about the database entries containing dynamic data referring to the respective static data point.

A. Dynamic and static data syncing

Ensuring data coherence relies heavily on the synchronization of data repositories. As previously mentioned, dynamic and static repositories exist concurrently, housing distinct datasets that may overlap in certain samples, such as sensor identifiers. Consequently, employing synchronization methods becomes a pivotal factor in upholding uniform identifiers across both repositories. The depicted approach can be observed in Fig. 6, illustrating two types of synchronization.

- The initial synchronization of the repositories, indicated by a dotted line, corresponds to the first loading of data into the repositories. This marks the commencement of the data population process.
- Second case lies in the dynamic syncing of data. When the ETL is executed, metadata is extracted and, then, sent to the static repositories to detect inconsistences, such as a new sensor has been installed.

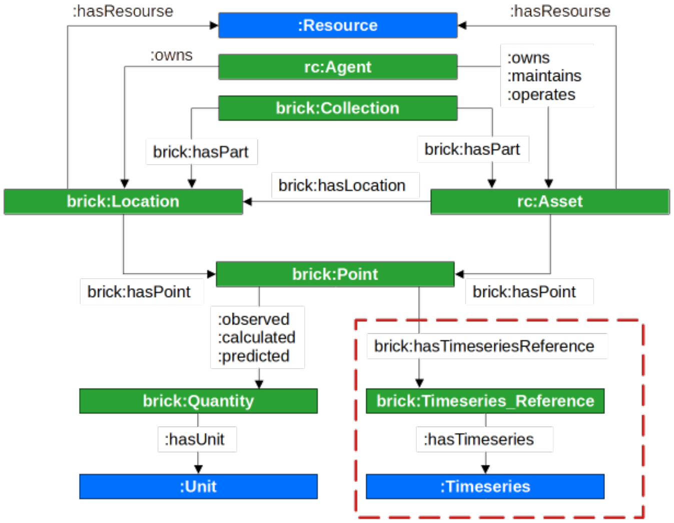

Fig. 5. UML diagram DigiBUILD's ontology core module

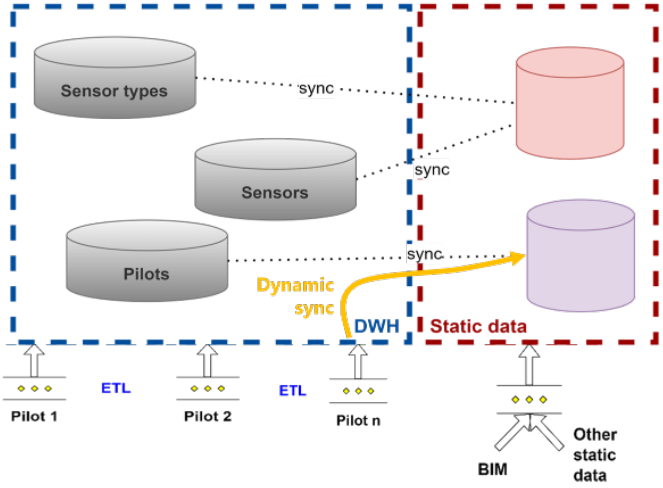

Fig. 6. Synchronisation of the static and dynamic repositories

V. Conclusions

The contemporary shift towards digitalisation is shaping a new era in building technologies, thanks to emerging digital tools. However, navigating this transformation is challenging, primarily due to the diverse array of data sources. The intricacies are compounded by the omission of crucial contextual data during the process. This necessitates the development of novel methodologies for data acquisition, aligned with innovative conceptual approaches to Extract, Transform, Load (ETL) procedures.

This research has explored ETL implementations across various pilots, revealing the intricacies of data gathering. Simultaneously, standardized mechanisms are introduced to harmonize data and augment timeseries information, facilitating better-informed decision-making. Yet, the ETL process alone falls short. Synchronisation between dynamic and static repositories becomes imperative to maintain data coherence. At times, the usability of data for smart services is hampered by inconsistencies between contextual information and timeseries data extracted from local databases. To address this, this work introduces a dynamic synchronization approach, ensuring continuous updates to graph databases and thereby enhancing data availability.

In upcoming endeavors, the fully implementation across all the pilots will open future implementations and research directions. Throughout this process, adaptations and new requirements may emerge, presenting opportunities to refine and enhance data enrichment strategies, making them more scalable across different architectures and following velocity, variety and volume of data.

Acknowledgment

The authors of this paper would like to thank DigiBUILD consortium, especially pilot leaders, involved in the project for the support and provision of interfaces to implement the ETLs.

References

[1] J. Hernandez, S. Martin, V. Marinakis, and I. de Miguel, "From silos to open, federated and enriched data lakes for smart building data management," in *2023 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*, pp. 29–33, 2023.

[2] J. L. Hernandez, S. Martin, P. Kapsalis, K. Katsigarakis, E. Sarmas, and V. Marinakis, "Building a data lake for smart building data: Architecture for data quality and interoperability," in *2023 14th International Conference on Information, Intelligence, Systems amp; Applications (IISA)*, (Los Alamitos, CA, USA), pp. 1–8, IEEE Computer Society, jul 2023.

[3] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim, "A review on optimized control systems for building energy and comfort management of smart sustainable buildings," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 409–429, 2014.

[4] European Commission, "Energy Performance Buildings Directive." https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en#current-rules-to-improve-the-eus-building-stock, 2023.

[5] 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, 2020.

[6] J. L. Hernández, R. Sanz, Corredera, R. Palomar, and I. Lacave, "A fuzzy-based building energy management system for energy efficiency," *Buildings*, vol. 8, no. 2, 2018.

[7] F. D. A. Pereira, C. Shaw, S. Martín-Toral, J. L. Hernández, R. S. Jimeno, D. Finn, and J. O'Donnell, "Towards semantic interoperability for demand-side management: a review of bim and bas ontologies," in *Proceedings of the 2022 European Conference on Computing in Construction*, vol. 3 of *Computing in Construction*, (Rhodes, Greece), European Council on Computing in Construction, July 2022.

[8] DigiBUILD, "High-Quality Data-Driven Services for a Digital Built Environment towards a Climate-Neutral Building Stock." https://digibuild-project.eu/, 2023.

[9] S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori, and M. Vincini, "A semantic approach to ETL technologies," *Data & Knowledge Engineering*, vol. 70, no. 8, pp. 717–731, 2011.

[10] K. Katsigarakis, G. N. Lilis, and D. Rovas, "A cloud IFC-based BIM platform for building energy performance simulation," in *EC3 Conference 2021*, vol. 2, pp. 350–357, 2021.

[11] P. S. Diouf, A. Boly, and S. Ndiaye, "Variety of data in the ETL processes in the cloud: State of the art," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pp. 1–5, IEEE, 2018.

[12] K. C. Mondal, N. Biswas, and S. Saha, "Role of machine learning in ETL automation," in *Proceedings of the 21st international conference on distributed computing and networking*, pp. 1–6, 2020.

[13] DigiBUILD, "DigiBUILD Ontology." https://ontology.digibuild-project.com/, 2023.