

DATA QUALITY IN THE BUILT ENVIRONMENT ACROSS EUROPE

José L. Hernández¹, Susana Martín¹, Vanja Lonac¹, Ignacio de Miguel² and Alessandro Rossi³

¹CARTIF technology centre, Boecillo, Spain

²Universidad of Valladolid, Valladolid, Spain

³Engineering Ingegneria Informatica S.p.A., Palermo, Italy

Abstract

Energy efficient projects are usually based on data to take better-informed decisions. High-quality data is necessary to reduce the uncertainty, which is a challenge to solve data gaps and/or outliers, thus limiting the trustworthiness in the decision-making. High-quality data is essential for assessing energy usage, identifying improvement measures and implementing effective solutions. This paper presents a data quality methodology, based on seven dimensions and validated in ten pilots across Europe. The maturity levels in the monitoring stage are crucial. Completeness, accuracy and consistency values vary between 40%-99%. It indicates the need for creating correction procedures and increasing the data quality.

Introduction

The increasing adoption of leading-edge Information and Communication Technologies (ICTs), such as Internet of Things (IoT), Artificial Intelligence (AI), Distributed Ledger Technology (DLT), Blockchain or Big Data (BD), as well as the wider extension of Building Automation and Control Networks (BACN) existence (mainly in tertiary buildings), has motivated the generation of big amounts of data in the building domain.

Even though over the past two decades, there has been an increasing focus on the digitalization of building data, and more and more data are being generated by the building stock, data quality remains a challenge (Duvier et al., 2008). There are data gaps, errors and inconsistencies in registrations that still need to be addressed. In contemporary construction and building projects, data plays a pivotal role in decision-making processes, project management, and overall success. However, the quality of data used in these endeavours often falls short of desired standards, posing significant challenges throughout project lifecycles. Poor data quality in building projects presents multifaceted obstacles, ranging from increased costs and delays to compromised safety and functionality.

Additionally, nowadays data-driven approaches are being followed, but here the main challenge is related to the reliability of the results due to data issues (e.g. communication or infrastructure problems) (Yong et al., 2021). One of the major challenges is to be able to process and analyse the data traceability to detect the errors (Hosseini et al., 2020). The big amounts of data being generated make this analysis more and more complex. In terms of data governance, there is no consensus about the meaning of this term (Ender, 2021). However, different types of data should be considered when determining the ownership, which is even more complex when non-human data is included. These data are produced and treated in

silos, without correlation between pillars (Abraham et al., 2019), which will promote better exploitation potential.

Poor data quality can also hinder effective stakeholder communication and collaboration, exacerbating misunderstandings and conflicts among project participants. In the absence of reliable and up-to-date information, decision-makers may resort to subjective judgments or outdated assumptions, further compounding project risks and uncertainties.

All these challenges are contributing to move forward towards the creation of a high-quality data-driven Smart Buildings Landscape. 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. Data-informed decision-making and digital upgrading can help unyield operational efficiencies at low cost.

A proper validation and detection of gaps, wrong and/or inaccurate data in the whole value chain of the building monitoring is a crucial step towards transparency and trust of the involved actors during the decision-making process to achieve the objectives of the Energy Performance of Buildings Directive (EPBD, 2023) of ensuring low usage of energy, low rate of carbon footprint, maximizing thermal comfort or assessing air quality.

According to the BDVA (Big-Data Value Association), the data quality concept is too absolute and misleading (BDVA/DAIRO, 2021). Within this paper, the high-quality data concept is based on the following aspects, establishing performance metrics as BDVA also suggests:

Reliability & Credibility: a data quality methodology is defined, deployed and applied, based on ML algorithms to learn about historical data, for reducing and correcting data errors not only in the data gathering process, but also in the propagation. This methodology takes care of:

- (1) Data gaps, in terms of missed data to detect non-completeness and creating interpolation methods when possible, to generate continuous data streams for dynamic timeseries. As well, for static and contextual data, completeness in terms of missing data to be completed with other sources (e.g. digital logbook missing data and cadastre information);
- (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-Building Information Modelling). 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. To solve this, dynamic and adaptive interoperability will be ensured by using standard Data Models (such as NSGI-LD, SAREF, BRICK or IFC, among others) to accommodate data into the requirements of the use cases or services to be deployed.

Privacy & Security: Last but not least, quality also involves privacy (personal data management) and security. DLT and blockchain framework can be applied here, where privacy and security are the main benefits.

Addressing the challenges stemming from poor data quality in building projects is paramount for the sustainable advancement of the construction industry. By assuring these aspects, buildings' stakeholders can rely on the data analytics and data-driven services thanks to high-quality data stocks. Additionally, the creation and use of data-driven business models will be built on top of high-quality data, providing more accurate and trustworthy value (e.g. ESCO-Energy Services Company models where energy prices are based on performance calculation and energy savings could be calculated using wrong data).

Real-world cases abound where the repercussions of inadequate data quality have been acutely felt, leading to suboptimal project outcomes. For instance, consider the scenario of a large-scale infrastructure development project where inaccurate site survey data resulted in improper foundation designs, ultimately leading to structural instability and substantial rework costs. Similarly, in the realm of building information modelling (BIM), incomplete or inconsistent data inputs have led to coordination errors, clashes between different trades, and inefficiencies in construction sequencing. Another example could be obtained from ESCO services, whose main challenge is the data collection from smart meters. Data gaps are usually appearing, leading to more complex billing procedures due to missing data, as well as the limitations in the energy savings calculation.

In summary, this paper presents how the previous challenges are overcome by deploying a data quality methodology that covers the data life-cycle of the building, delving into the complexities surrounding data quality issues. It will not be only focused on the data gathering at field level, but it will cover the different stages of the building and data lifecycle to reduce the error propagation. This methodology will be supported by a federated Data Lake (Hernández et al., 2023a) (deployed within a data platform) to collect cross-domain data, being

able to benefit the data management and governance processes. Finally, the data platform will provide additional services, which will facilitate the management of data ownership, and together with blockchain, will allow traceability, privacy and security of data stocks.

The rest of the paper is structured as follows. The following section presents the literature review in terms of data quality (including real-case scenarios). Next, the data quality methodology is explained, to be later applied in 10 pilots composed by one or several buildings. Results of the data quality assessment are collected into the "Data quality analysis results" section, where statistics for the 7 data quality dimensions are presented per pilot. A "Discussion" section follows, to highlight the dependency of the buildings' data quality values to the age of the building and monitoring maturity and stage. Finally, the main remarks of the paper are presented in the "Conclusions" section.

Literature review

The literature provides various definitions and dimensions of data quality in the context of smart buildings. While traditional data quality dimensions such as accuracy, completeness, consistency, and timeliness remain relevant, the dynamic nature of IoT-generated data introduces new challenges and considerations. The importance of additional dimensions such as interoperability, security, privacy, and relevance in assessing data quality in smart buildings is emphasized (Rao, 2023). Ensuring data compatibility and integration across heterogeneous systems while safeguarding privacy and security concerns emerges as critical areas of focus.

Despite the potential benefits, smart building deployments encounter numerous challenges related to data quality. These include issues such as data silos (Hernández et al., 2023b), interoperability gaps between devices and systems (Coujard, 2023), sensor inaccuracies, data integration complexities, and cybersecurity vulnerabilities. The literature highlights organizational barriers such as lack of data governance frameworks (Kaginalkar, 2023), limited expertise in data management, and resistance to change as significant impediments to achieving and maintaining high data quality standards in smart building initiatives.

Diagnosis of data quality is therefore a main challenge. Some authors have analysed the uncertainty in the calibration of the measurements (Morewood, 2023). It is reported that, in 43 out of 63 real-cases, overall accuracy is achieved, while precision is reduced up to four. Calibration is shown in 18, whereas measurement out-of-range is only reached in 23. Another real-case scenario is performed in the cities of Nantes, Hamburg and Helsinki (Hernández et al., 2022). The main conclusion is the real need for methods to increase data quality to foster better-informed decisions. The authors demonstrated the low quality of various data-sets in terms of completeness and out-of-range, highlighting the importance of proper commissioning mechanisms.

Complementary to these previous studies, the contribution of this work extends the data quality analysis to additional real-case sites across Europe to understand the real data-quality scenarios. Moreover, while previous researches limited the dimensions of the data quality, this manuscript increases the dimensions, covering accuracy, completeness, reliability, consistency, relevance, accessibility and timeliness. This work has been carried out within the EU HORIZON DigiBUILD project, whose main aim is to transform traditional silo approaches by making use of high-quality data and next generation digital building services for assuring trust and transparency, and better-informed decision-making processes. It counts on ten pilots across Europe where the technologies are tested.

Data quality methodology

Within DigiBUILD project, a multi-dimension strategy is approached in terms of data quality. Table 1 summarizes the list of dimensions that are part of the data quality analysis. From the 7 dimensions, accuracy, completeness and consistency are considered the pivotal ones.

Table 1: Data quality multi-dimension approach

Data quality dimension	DigiBUILD approach
Accuracy	Out of “expected” range detection
Completeness	Gap detection and interpolation
Reliability	Accuracy x completeness
Consistency	Detection of “normal” data patterns & Proof of Existence (PoE)
Relevance	Data filtering processes
Accessibility	Open APIs and protocols
Timeliness	Pre-analytics

Accuracy determines the percentage of data samples compliant with the “expected” ranges. It should be clarified the “expected” range differs from the measurement range. For instance, an indoor temperature sensor can obtain data within values from 0°C to 50°C (depending on the manufacturer). However, temperature are expected to be around 15-30°C (depending on the building use, insulation level and other features). The accuracy dimension treats the detection of data samples, which, even correct, are out of this range to avoid its usage in the application of services (i.e., classified as outliers). Within DigiBUILD, the procedure calculates the data samples within the range specified by the building expert, manager and/or operator to extract the ratio with respect to the total number of samples.

Completeness goal is to determine the data gaps that are in a dataset. Normally, data samples are periodically collected according a sampling frequency. Therefore, the total samples to be gathered in a time span are known.

Thus, the missing data can be calculated by the ratio between the available data samples and the theoretical ones to be saved.

Reliability focuses on the feasibility of data to carry out analysis, develop high-level services or provide high-quality data to third parties. This dimension correlates the completeness and accuracy to provide an overall mark.

Consistency looks for the data patterns within data. For that end, normal distribution is considered, where measurements are scattered between the mean and \pm standard deviation.

Relevance dimension is helpful to determine the datasets that are really useful from the end-user’s perspective (services). Data normally include data-points without any usability in the implementation of the services; therefore, the percentage of relevant data is the main purpose.

Accessibility provides the ability to access, gather and share data. Here, the use of open protocols and APIs (Application Programming Interfaces) fosters the accessibility to the datasets. Nevertheless, this is not always possible, limiting the access to some datasets. Thus, this dimension checks the ability to access the relevant data from the building.

Timeliness is related to up-to-date data or the delay to gather the information that is required for making informed decisions. It is expressed in time units in order to establish the maximum delay when data is made available.

Table 2: Data quality outcomes within DigiBUILD

Data quality dimension	DigiBUILD outcome
Accuracy	> 85%
Completeness	> 90%
Reliability	No duplicates (clean data)
Consistency	Increase 10%
Relevance	90% of useful datasets
Accessibility	90% of data accessible
Timeliness	Specific per pilot

With regard to the results to be obtained in DigiBUILD, Table 2 shows the outcome that should be achieved by the detection and the correction techniques to be applied. In this sense, the accuracy and completeness dimensions should reach 85% and 90% respectively to provide a clean dataset without duplicates (reliability). Moreover, the case of consistency aims to improve the quality of the datasets being used in a 10%, making relevant the 90% of the collected dataset with an accessibility of 90%. Finally, timeliness depends on the specifications of the pilots when data is polled / pushed.

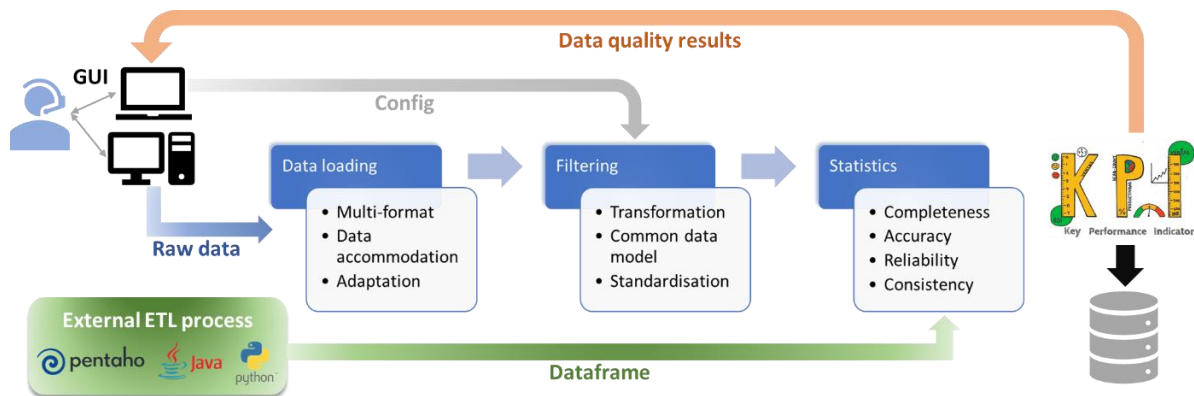


Figure 1: Data quality detection and check approach

Data quality check approach

Two procedures are approached for data quality assurance. On the one hand, the checker in charge of the analysis of the quality of datasets; on the other hand, the methodology for the correction of data (such as interpolation or other techniques for data cleansing). However, this paper is focused on the status of the data quality across Europe; therefore, the correction part is out of the scope. It will consist of a machine-learning approach to learn from historical data. Then, when a gap or an inconsistent data measurement is obtained, the corrector module will populate data according to the data-driven model.

Focusing on the checker, Figure 1 depicts the schema how it is approached within DigiBUILD. It works on two possibilities. The first one makes use of downloaded data from the pilots, for instance, csv files, Excel sheets, JSON documents, etc. This raw data is ingested by a data loading module that accommodates the multiple formats and adapts them in order to be transformed (filtering process) into a common data model. The goal is to provide a “standard” representation of data (following a python DataFrame). Complementary, a configuration file is used where the user provides the parameters and specific aspects related to data.

The second way is the use of data collection technologies, as for the case of Pentaho (Hitachi, 2023), python or Java scripts. In this case, the tools for data interfacing already provide the data loading and filtering processes. Hence, they are directly connected to the statistics module.

The statistics module is a common component working in the two paths. It basically calculates the indicators for the data quality in the multiple dimensions explained before. It accepts a DataFrame in order to make it interoperable, replicable and scalable; therefore, being able to calculate the data quality indicators independently of the source.

Data quality analysis results

As introduced, this paper presents the baseline results in terms of data quality for a set of buildings (pilots) across Europe. As extracted in Table 3, ten pilots from different European countries that are part of the EU DigiBUILD

project, with different typologies and available datasets are analysed. Next sections describe the results for each one of these pilots in the various dimensions that are accounted in DigiBUILD.

Table 3: Pilot buildings within the EU DigiBUILD project for the data analysis

Pilot	Type	Datasets
UCL (UK)	University	Energy, photovoltaics, indoor air quality, occupancy, windows contact
EDF (France)	Offices	Indoor air quality, energy (heating, electricity and plugs)
IASI (Romania)	Various types	Indoor air quality, energy (heating and electricity)
VEOLIA (Spain)	Residential	Energy (district heating)
EMOT (Italy)	Offices	Energy, photovoltaics, charging station, electric vehicle, indoor air quality
FOCCHI (Italy)	Factory	Energy, photovoltaics, indoor air quality
HERON (Greece)	Various types	Energy, charging station, electrical vehicle
FVH (Finland)	Various types	Energy (district heating & heat pumps), photovoltaics
IEECP (The Netherlands)	Schools	Energy, indoor air quality, occupancy
NTUA (Greece)	University	Energy, indoor air quality, HVAC info

Accuracy

Starting with the accuracy dimension, The results show a dispersion of values, from 42.04% to 99%. In some cases, as for instance the UCL pilot, out- of-value ranges are due to changing needs of the heating/cooling systems. Other cases, malfunctioning of any sensor causes measurements out of range.

Table 4 summarizes the percentages obtained from a historical dataset sample of the aforementioned available datasets in the pilot building. For that end, minimum and maximum values for each data-point are obtained in order to reduce the range in 10%, being an acceptable value,

such as stated by (Villada et al., 2008) and (Seyedzadeh et al., 2020). Then, values out of these range (known as expected) are considered anomalous.

The results show a dispersion of values, from 42.04% to 99%. In some cases, as for instance the UCL pilot, out-of-value ranges are due to changing needs of the heating/cooling systems. Other cases, malfunctioning of any sensor causes measurements out of range.

Table 4: Accuracy results of the data analysis

Pilot	Accuracy
UCL (UK)	70%
EDF (France)	44.43%
IASI (Romania)	n.a.
VEOLIA (Spain)	85%
EMOT (Italy)	56.18%
FOCCHI (Italy)	69.92%
HERON (Greece)	42.04%
FVH (Finland)	99%
IEECP (The Netherlands)	86.90%
NTUA (Greece)	55.69%

Completeness

Completeness aims the detection of data gaps in the timeseries. There are many cases where communication errors, malfunctioning of sensors or processing errors, among others, provoke data logging problems. Table 5 highlights the results of the analysis per pilot.

The results are diverse, remarking the case of EMOT, which is offering very low values (approximately 62% of data samples are missing). It should be considered the case of VEOLIA, with very high values, although this is not the reality. The reason behind lies in interpolation from the heat meter in a form of linear regression. The appearance is very complete data-set, but, with accurate populated values (e.g., boilers not working, but linear interpolation indicating energy consumption).



Figure 2: Monitoring stages

It is notable the maturity levels in the data collection. Pilots such as EDF or FVH with multiple years of data collection offers higher values with a very stable data gathering procedure. Less mature pilots are still in early stages where optimal operation is still not reached, as depicted in the monitoring stages in Figure 2 (Hernández et al., 2022).

Table 5: Completeness results of the data analysis

Pilot	Completeness
UCL (UK)	70%
EDF (France)	99.82%
IASI (Romania)	n.a.
VEOLIA (Spain)	99%
EMOT (Italy)	38.39%
FOCCHI (Italy)	~100%
HERON (Greece)	67.31%
FVH (Finland)	75.24%
IEECP (The Netherlands)	98.78%
NTUA (Greece)	99%

Reliability

The case of reliability focuses on the error-free data, without inconsistencies (accurate and complete). Then, its value is represented by the correlation between the completeness and accuracy, considering unique values from completeness (i.e., no data duplicates). Table 6 provides the results, where it is worth mentioning the case of EDF, in contrast to the completeness, lots of duplicates are removed, decreasing the reliability value.

Table 6: Reliability results of the data analysis

Pilot	Reliability
UCL (UK)	50%
EDF (France)	44.35
IASI (Romania)	n.a.
VEOLIA (Spain)	85%
EMOT (Italy)	21.57%
FOCCHI (Italy)	62.92%
HERON (Greece)	28.29%
FVH (Finland)	74.49%
IEECP (The Netherlands)	85.84%
NTUA (Greece)	55.63%

Consistency

Consistency relates to uniformity of data in terms of data behaviour. Considering data distribution follows a “normal” curve around the mean value, it is considered

that data is consistent when a data-point complies with mean \pm standard deviation. Table 7 illustrates the values for consistency. It should be taken the unexpected behaviour of the systems into account, such as the heating system. An example could be understood with COVID-19 situation, where energy demand increased, being a non-usual operation. Therefore, producing a deviation with respect to the “normal” distribution.

Table 7: Consistency results of the data analysis

Pilot	Consistency
UCL (UK)	70%
EDF (France)	34.11%
IASI (Romania)	n.a.
VEOLIA (Spain)	79%
EMOT (Italy)	32.97%
FOCCHI (Italy)	71.83%
HERON (Greece)	88.60%
FVH (Finland)	~99%
IEECP (The Netherlands)	82.14%
NTUA (Greece)	86.53%

Consistency is also addressed by a Smart Contract in charge of the notarization of data. A Proof of Existence (PoE) algorithm is developed, based on blockchain, to calculate the hash on the data provider side, as well as on the data storage side (see Figure 3). By comparing the hashes, the consistency in both sides can be obtained (i.e., detection of communication errors or data manipulation).

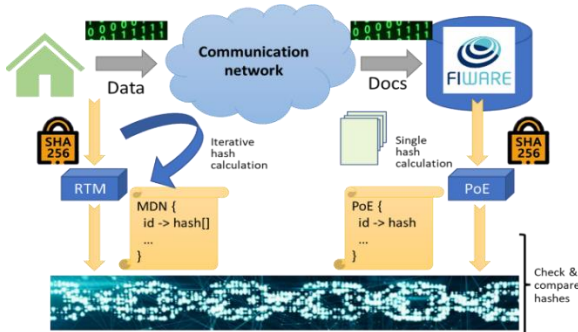


Figure 3: Proof of Existence concept

Figure 4 depicts the process, which has been applied within VEOLIA case. From the real data read, a data-set is filtered and loaded as a DataFrame, whose hash is calculated to be sent to the blockchain as a transaction and, thus, certify data. In the verification step, the received data is hashed and compared to the original hash (see Figure 4). The result of the transaction is shown below, where the output from the blockchain is generated,

indicating the hash of the transaction and the block that has been generated.

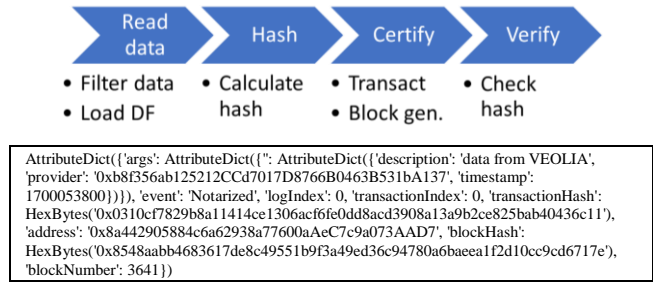


Figure 4: Consistency check from Smart Contract

Relevance

The relevance dimension focuses on the end-user and the usability of data. Nowadays, monitoring systems and digitalization have exponentially grown, obtaining bigger amounts of data without real usability in services or decision-making. That is why relevance plays an important role in DigiBUILD, determining the real usable datasets to provide high-level services. Table 8 specifies the percentage of datasets that are relevant for the DigiBUILD services and, thus, useful in this context.

Table 8: Relevance results of the data analysis

Pilot	Relevance
UCL (UK)	35.54%
EDF (France)	28%
IASI (Romania)	37.5%
VEOLIA (Spain)	55.73%
EMOT (Italy)	100%
FOCCHI (Italy)	55.73%
HERON (Greece)	100%
FVH (Finland)	30%
IEECP (The Netherlands)	n.a.
NTUA (Greece)	n.a.

Accessibility

Data, even with quality enough and relevant, sometimes is not accessible due to proprietary protocols. Accessibility determines the ability of remotely capture data and make it available for services and decision-making procedures. In this sense, Table 9 compiles the percentage of datasets that are accessible from pilot buildings. In some cases, although interfaces for access are being configured, these are not fully available; therefore, limiting the accessibility to data.

Table 9: Accessibility results of the data analysis

Pilot	Accessibility
UCL (UK)	50%
EDF (France)	50%
IASI (Romania)	n.a.
VEOLIA (Spain)	100%
EMOT (Italy)	80%
FOCCHI (Italy)	50%
HERON (Greece)	80%
FVH (Finland)	80%
IEECP (The Netherlands)	70%
NTUA (Greece)	80%

Timeliness

Last dimension in DigiBUILD is timeliness, responsible for setting the time when data is made available, which relates to the “real-time” operation. Some samples are obtained in 5 min periodicity, while others are in 15 minutes or even 1 hour, as specified in Table 10. Here, the event-driven data collection (in other words, change of value) is not considered, although it exists.

Discussion

Data quality is crucial to make decisions, such as improvement in the energy management of buildings, retrofitting investments or air quality enhancement. This is not always covered and decisions are made based on low quality data. Within DigiBUILD a methodology to check datasets quality, incorporating the improvement, is proposed. At this first stage, the initial data analysis has been performed to determine the baseline.

What is clearly remarkable is the maturity level and the age of the building. According to Figure 2, depending on the monitoring stage, optimal operation in the data collection can be achieved, as the cases of VEOLIA, IEECP or FVH pilots, with very high level of maturity; therefore, high-quality of data. Others, such as IASI pilot is still in the interventions stage, without providing data (0% in the indicators), as depicted in Figure 5. The age of the building is also important. New buildings, such as the UCL pilot with only 1 year old, contain smarter technology with capabilities of data storage and access.

To sum up, it is undoubtedly data quality is a pivotal aspect in the decision-making process and this aspect should be carefully treated. As observed in Figure 5, there exists a diversity of maturity levels with a clear room for improvement. The graph is only showing a subset of dimensions (the most relevant ones). Availability and

accessibility of data must be ensured. Data processing for data filtering and cleansing is a very important task that would improve the current data quality.

Table 10: Timeliness results of the data analysis

Pilot	Timeliness
UCL (UK)	5/15 min
EDF (France)	5/10 min
IASI (Romania)	n.a.
VEOLIA (Spain)	15 min
EMOT (Italy)	3 min
FOCCHI (Italy)	1 hour
HERON (Greece)	30 sec
FVH (Finland)	10 min
IEECP (The Netherlands)	30 min
NTUA (Greece)	5 min

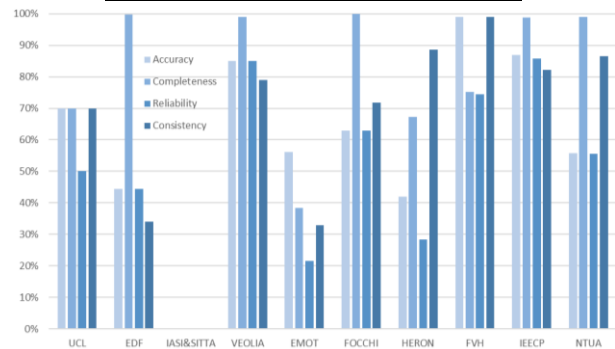


Figure 5: Data quality indicators per pilot building

Conclusions

This paper has described a data analysis carried out under the umbrella of the European DigiBUILD project to determine the baseline in terms of data quality for decision-making. From historical data samples in 9 different countries across Europe, 7 dimensions have been considered within the data analysis. The results provide an overview of the real maturity levels within the digitalization and data collection of the building stock.

Different levels of maturity could be observed, from pilots that already are in the optimal operation stage; then, obtaining high-values for the different data quality dimensions to other in very early stages, decreasing the trustworthiness on data. This deals with two issues. First of all, making decisions based on low-quality data; hence, increasing the uncertainty. Secondly, decision-makers are reluctant when addressing an energy efficiency project. In this sense, data quality checks are crucial.

A data analysis has been performed to check the data quality levels of multiple datasets. Due to the capability

for improvement of the quality, the methodology is extended with correction techniques. At the current state of the project, they have not been fully implemented and it is the future work. The aim is to increase data quality and reach the outcomes that were previously explained.

Acknowledgments

The authors would like to thank DigiBUILD consortium and European Commission for funding the project under GA#101069658 of the Horizon Europe programme.

References

- Abraham, R., Schneider, J., Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 49, pp. 424-438.
- BDVA/DAIRO position paper. (2021). Response to the European Commission's proposal for AI Regulation. https://www.bdva.eu/sites/default/files/BDVA_DAIRO%20response-feedback%20AI%20Regulation_Final.pdf, visited on 10th January 2024.
- Coujard, C., Eloire, K. L., Zarli, A., & David, A. (2023). SmartBuilt4EU: Towards a strategic research and policy agenda for the European smart buildings community. In *ECPPM 2022-eWork and eBusiness in Architecture, Engineering and Construction 2022* (pp. 785-794). CRC Press.
- Duvier, C., Neagu, D., Oltean-Dumbrava, C., Dickens, D. Data quality challenges in the UK social housing sector, *International Journal of Information Management*, Volume 38, Issue 1, 2018, 196-200.
- Ender, L. 'Data Governance in Digital Platforms: A case analysis in the building sector', Dissertation, 2021
- EPBD (2023) https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en#current-rules-to-improve-the-eus-building-stock
- Kaginalkar, A., Kumar, S., Gargava, P., & Niyogi, D. (2023). Stakeholder analysis for designing an urban air quality data governance ecosystem in smart cities. *Urban Climate*, 48, 101403.
- Hernández, J.; Quijano, A.; García, R.; Nouaille, P.; Risch, L.; Virtanen, M. and de Miguel, I. (2022). Analysis of Data Quality in Digital Smart Cities: The Cases of Nantes, Hamburg and Helsinki. In *Proceedings of the 11th International Conference on Data Science, Technology and Applications - DATA*; ISBN 978-989-758-583-8; ISSN 2184-285X, SciTePress, pages 353-360. DOI: 10.5220/0011271900003269
- Hernández, J., Martín, S., Kapsalis, P., Katsigarakis, K., Sarmas, E., Marinakis, V. Building a Data Lake for Smart Building Data: Architecture for data quality and interoperability, 2023 14th International Conference on Information, Intelligence, Systems and Applications (2023a), University of Thessaly, Volos, Greece, 10-12 July 2023.
- Hernández, J. L., Martín, S., Marinakis, V., & de Miguel, I. (2023b). 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). IEEE.
- Hitachi, Pentaho Data Integration, online: https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.1/Products/Pentaho_Data_Integration, visited on 27th December 2023.
- Hossein Motlagh, N.; Mohammadrezaei, M.; Hunt, J.; Zakeri, B. Internet of Things (IoT) and the Energy Sector. *Energies* 2020, 13, 494. <https://doi.org/10.3390/en13020494>.
- Morewood, J. (2023). Building energy performance monitoring through the lens of data quality: A review, *Energy and Buildings*, Volume 279, 2023, 112701, ISSN 0378-7788, <https://doi.org/10.1016/j.enbuild.2022.112701>.
- Rao, P. M., & Deebak, B. D. (2023). Security and privacy issues in smart cities/industries: technologies, applications, and challenges. *Journal of Ambient Intelligence and Humanized Computing*, 14(8), 10517-10553.
- Seyedzadeh, Saleh, Rahimian, Farzad Pour, Oliver, Stephen, Rodriguez, Sergio and Glesk, Ivan, Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making, *Applied Energy*, Volume 279, 2020, 115908, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2020.115908>.
- Villada, Fernando; Cadavid, Diego Raúl and Molina, Juan David. Pronóstico del precio de la energía eléctrica usando redes neuronales artificiales. *Rev.fac.ing.univ. Antioquia* [online]. 2008, n.44 [cited 2023-12-27], pp.111-118. Available from: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-62302008000200011&lng=en&nrm=iso. ISSN 0120-6230.
- Yong Teng, S., Touš, M., Dong Leong, W., Shen How, B., Loong Lam, H., Máša, V. Recent advances on industrial data-driven energy savings: Digital twins and infrastructures, *Renewable and Sustainable Energy Reviews*, Volume 135, 2021, 110208, <https://doi.org/10.1016/j.rser.2020.110208>