# Enhancing Data Quality in Large-Scale Software Systems for Industrial Automation

Valentina Golendukhina
University of Innsbruck
Innsbruck, Austria
valentina.golendukhina@uibk.ac.at

Lisa Sonnleithner
CDL VaSiCS, LIT CPS Lab
Johannes Kepler University Linz
Linz, Austria
lisa.sonnleithner@jku.at

Michael Felderer
German Aerospace Center (DLR)
Cologne, Germany
University of Innsbruck
Innsbruck, Austria
michael.felderer@dlr.de

## ABSTRACT

Modern industrial systems have become highly automated and data-driven, generating large volumes of data through sophisticated machinery. However, the quality of the collected data is not always optimal, whereas monitoring data quality is challenging due to real-time data constraints. While significant research has been done on data validation of the exported and prepared data, there is no research on implementing data quality practices with programming languages and tools that directly interact with hardware in the domain of cyber-physical production systems (CPPSs), such as IEC 61499 and IEC 61131-3, i.e., software on level 1 of the automation pyramid. By examining a plant-building company, this short paper explores the challenges and opportunities for data quality management at L1 including knowledge transfer, data compression, and metadata formulation, and suggests possible data validation techniques.

## CCS CONCEPTS

• **Software and its engineering → Software verification and validation**.

## KEYWORDS

data quality, cyber-physical production system, IEC 61499

**ACM Reference Format:**
Valentina Golendukhina, Lisa Sonnleithner, and Michael Felderer. 2023. Enhancing Data Quality in Large-Scale Software Systems for Industrial Automation. In *Proceedings of the 3rd International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things (SEA4DQ '23), December 4, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3617573.3618028

## 1 INTRODUCTION

Modern industrial systems have undergone a significant transformation, becoming highly automated and data-driven. Equipped with an extensive array of sensors and sophisticated machinery, these systems generate and accumulate vast amounts of data. Such data plays a critical role in facilitating decision-making processes, ensuring safety and security, and optimizing operational efficiency. However, the quality of the collected data is not always optimal, posing challenges for effective utilization. Hardware issues, proximity to magnetic fields, and incorrect settings can potentially corrupt the data [12]. Furthermore, maintaining and monitoring data quality is challenging due to the large amount of data exchanged close to real-time and the changing and noisy operational environments in which these systems are typically deployed [10].

According to the automation pyramid, data is collected and managed at four levels [1]. The initial level (L1) involves the software that directly interacts with the hardware configurations of sensors within a system via programmable logic controllers (PLCs). The higher processing level (L2) represents a more advanced system equipped with enhanced computational capabilities to handle data monitoring and supervision. The third and fourth levels (L3 and L4) focus on data management, optimization, and business planning.

Significant research efforts are dedicated to data validation at L3 and L4, some practices are implemented at L2. These stages have more computational power to ensure the accuracy, consistency, and reliability of the processed data and enable various techniques including machine learning-based data mining techniques, statistical outlier detection methods, and various mathematical models [2]. However, addressing certain issues at L1 can potentially contribute to improving data quality, but was not investigated in detail. By implementing effective strategies at this level, the amount of unnecessary or redundant data transmitted can be minimized. This, in turn, results in substantial savings in terms of computational and storage capacities, as well as reducing the workload for data analysts involved in data cleaning tasks.

This paper explores the challenges associated with data quality in large industrial systems that can be potentially addressed on L1, focusing on the framework provided by the International Electrotechnical Commission (IEC) 61499/61131-3 [6]. IEC 61131-3 defines programming languages for PLCs and is widely spread in the domain of industrial automation. IEC 61499 extends IEC 61131-3 to provide means for modeling distributed control systems. By examining the current challenges and opportunities for data validation, we aim to enhance the understanding of data quality management at L1 and propose potential solutions within the context of large-scale industrial systems.

The remaining sections of the paper are structured as follows. Section 2 describes the IEC standard, its application, and research on data quality in the industrial domain. Section 3 presents the research setting and discovered data quality challenges. In Section 4, we discuss the implications for data quality improvement. Finally, Section 5 concludes the paper with the future work.



*SEA4DQ '23, December 4, 2023, San Francisco, CA, USA*

ACM ISBN 979-8-4007-0378-2/23/12.
https://doi.org/10.1145/3617573.3618028

# 2 BACKGROUND

## 2.1 IEC 61131-3 and IEC 61499

IEC 61131-3 [5] is a standard that describes five languages for PLC programming. Instruction List (IL) is a textual language that is similar to assembly. Structured Text (ST) is also a textual language that is syntactically similar to Pascal. Ladder Diagram (LD) is a graphical language that resembles a circuit diagram. Sequential Function Chart (SFC) is another graphical language that is similar to a state machine. The fifth language, Function Block Diagram (FBD), is a graphical language too. The Continuous Function Chart (CFC), is not officially part of IEC 61131-3. It is an adapted version of the FBD. The Function Blocks can be programmed in any language defined in IEC 61131-3.

IEC 61499 [6] defines a domain-specific modeling language for distributed control systems. It extends IEC 61131-3 and provides improved encapsulation of the software for increased reusability of software components. It also provides a vendor-independent format to improve portability and interoperability. The language is block-based and follows an event-based execution order. Function Blocks (FBs) can contain a state machine, can be programmed in any textual language (usually ST) or encapsulate other FBs.

PLCs are mainly programmed with the languages defined by IEC 61131-3, and IEC 61499 is in the phase of early adopters [8]. In the context of the automation pyramid, PLC software is on Level 1 (L1). It directly interacts with the hardware (e.g., sensors that send data) and forwards certain information to higher level software on L2, see Figure 1.

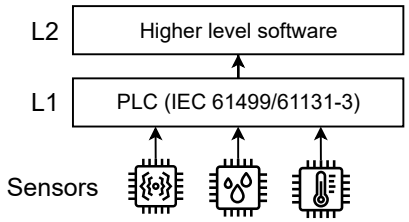


**Figure 1: Levels 1 and 2 of automation pyramid**

## 2.2 Data Quality

Since data quality is a crucial factor for the correct functioning of modern systems, a lot of work has been done to address data quality issues. Alwan et al. linked all data quality challenges in large-scale cyber-physical systems to three root causes: incorrect sensors' measurements, hardware failures of sensors or communication networks, and mismatches in spatial and temporal attributes [2].

There is a large number of methodologies available in the literature for evaluating and enhancing data quality focusing on technical aspects [3]. Some approaches concentrate on statistically assessing the quality of data, employing statistical and machine learning techniques to identify anomalies or inconsistencies within datasets [11]. Other approaches prioritize the evaluation of data sources themselves, aiming to predict the quality of the resulting data based on characteristics of the sources [4] or evaluate the data based on data provenance techniques by analyzing every step of data generation and transformation [7].

All described techniques are currently applied in the data analysis at L2 and higher. To the best of our knowledge, there is no research done to understand the applicability and application potential of data quality techniques at L1.

# 3 CASE STUDY

To analyze the current state of practice regarding data quality practices implemented on L1, we conducted two semi-structured interviews with two experts (a senior developer and a senior team lead, both with more than 10 years of experience in the domain) from an internationally operating large plant-building company. The interviews were held online and took 1 - 1.5 hour each. During the interviews, we explored the following topics:

- Description of sensors
- Description of the raw data from sensors on L1
- Description of data processing procedures on L1
- Description of the data sent to L2
- Current problems with sensors
- The procedure of data issues detection on L1 and L2

As a result, we could understand the company's setting and explore the challenges they face in practice. Following the thematic content analysis method, we distinguished three groups of challenges that are described in the following sections.

## 3.1 Industrial Setting

The company mainly uses CFC (block-based, graphical) and ST (textual) to program their control software. They utilize a combination of digital and analog sensors to monitor and control various aspects of a plant's equipment and processes.

For digital sensors, the company employs limit switches to detect the movement and position of equipment. These switches can be mechanical or electronic, such as inductive sensors commonly used in plants. They provide binary signals (0 or 1) and are primarily used to stop machine movement or indicate specific positions. Quality control is not integrated into these limit switches, so a broken wire would result in a 0 signal. Additionally, fail-safe limit switches are utilized in restricted areas, such as safe locks or tools. These sensors continuously check for wire breaks and promptly switch off the equipment when detected.

In addition to digital sensors, the company employs analog sensors to measure temperature, pressure, and gas flow. These sensors provide results typically within a 4 mA to 20 mA range. If the measurement falls below 4 mA, it indicates a sensor fault or a wire break. An external sensor vendor organization defines the thresholds for normal and abnormal sensor functioning.

After receiving the raw values from the sensors, the data is sent to a programmable logic controller (PLC) for further processing. To convert the raw values into meaningful process values, a special function block is employed. This function block applies transformations based on predefined mappings or calculations, converting the raw value into units such as bars, millibars, or degrees, depending on the specific measurement parameters. The range for these process values is determined in collaboration with domain experts.

## 3.2 Challenges

We identified three groups of data quality-related challenges that could be potentially addressed on L1.

*3.2.1 Data Volume.* A complete plant generates approximately 10,000 information values that need to be stored at a high frequency, up to every 4 to 10 milliseconds. Handling such a massive influx of data presents significant data compression and storage challenges.

When transmitting the data to L2, data compression becomes a crucial step to manage the immense data volume efficiently. L2 focuses on identifying relevant information such as peak values and critical data points during plant startup and shutdown, rather than continuous fluctuations. By compressing the data, they aim to save storage space while retaining critical data points relevant to these events. However, this compression strategy introduces the challenge of initially sending all data points to L2, demanding significant computational and storage resources for adequate data transportation and processing.

At L1, several functions are suitable to gather and analyze data, including functions like minimum, maximum, and average calculations. They could enable quick identification of peaks and search of crucial data points. However, compression functions are not directly employed on L1 due to concerns about potential data loss. Compression techniques adapted to L1 that ensure minimal data loss while achieving significant storage savings is of interest within L1 data management. This could optimize data quality while addressing storage and network limitations in modern industrial systems.

*3.2.2 Metadata Formulation.* Each signal gathered at Level 1 (L1) of the automation pyramid comprises diverse metadata. Despite the importance of it for data interpretation, the effectiveness of metadata practices at L1 varies.

The whole automation system is time synchronized enabling access to historical data of each sensor for up to one or two years based on project specifications. Additionally, deviations and predefined events are recorded in the alarm system. Each alarm occurrence is stored with a corresponding timestamp and description. However, one limitation arises from the absence of information about the quality of time synchronization. In cases where different sensors become asynchronous, accurately discerning the precise events and their sequence becomes challenging, potentially impacting the ability to thoroughly analyze and understand the causes and effects of alarms and achieve the best optimization.

While date-time synchronization continues to be a significant challenge in industrial data analysis, in certain sectors, where precise timing is of paramount importance, more robust approaches to time synchronization are adopted but it is not a standard practice.

Another possible problem of metadata management is that the data structure does not inherently store measurement units, necessitating the need to maintain this information externally.

To mitigate this issue, the measurement units are defined and assigned at the project's launch and then displayed in HMI (human-machine interface) at L2 indicating the appropriate unit. Documenting this information is essential since different clients use different measurement units. For instance, the measurement unit for pressure is standardized in the International System of Units (SI) as pascal (Pa) but European clients may prefer measurements in the non-SI unit bar ($1\,\text{Pa} = 10^{-5}$ bar), while US projects normally utilize PSI (pounds per square inch) ($1\,\text{Pa} = 1.4504 \cdot 10^{-4}$ PSI). However, there is no indication of measurement units at L1 which could lead to the inability to read the unit and the process value directly from sensors and other unexpected issues such as misinterpretation of values or insufficient documentation at L1.

Another current issue is the transparency of substitute values. If a sensor or a wire breaks, the need to maintain continuous process operations requires the use of substitute values. These substitute values are temporary placeholders to keep the process running smoothly until the sensor or wire is repaired or replaced. To ensure that these substitute values are properly handled and considered in the data management process, it is crucial to communicate their status and significance clearly between L1 and L2.

At the L1 level, when a substitute value is activated, it is essential to include clear metadata indicating its status and the reason for its use. This information can be stored and documented within the data structure itself. By doing so, the L2 process model and HMI systems can be made aware of the substitution and interpret the data correctly. Otherwise, L2 may treat the substitute values as normal process data leading to potential issues in data analysis.

*3.2.3 Knowledge Transfer.* Problems with knowledge transfer can further compound the challenges in data quality within large-scale software systems for industrial automation. As projects evolve over time, teams may change, and valuable knowledge about the data structure, metadata conventions, and best practices can be lost. Additionally, new team members may overlook important metadata, or misinterpreting the significance of certain values, including substitute values. Global knowledge transfer is an especially urgent problem in the domain of industrial systems, where the vendors provide and install the systems, but the runtime maintenance in the next years is the responsibility of customers.

Understanding how customers conduct maintenance work, the frequency of clean-ups, and the planning process would be beneficial for creating an effective diagnostic system that can learn from their experiences and provide valuable insights. Potentially, this knowledge could be incorporated into L2 for process optimization or as an add-on to L1 and could bridge the gap between process optimization and control. By having access to detailed maintenance data, the system could offer tailored recommendations, predictive maintenance schedules, and performance improvements, resulting in more efficient operations and reduced downtime.

However, customers are often unwilling to share such information because of concerns about the protection of their proprietary knowledge. As a result, there is a delicate balance between providing valuable insights to customers while encouraging them to contribute feedback to enhance the system further.

## 4 DISCUSSION

In the industrial domain, the focus on communication protocols and interoperability often leaves a gap when it comes to explicitly addressing data quality requirements. Monostori describes several challenges of CPPS including operating sensor networks, handling big bulks of data, information retrieval, representation, and interpretation, and security aspects [9]. Notably, a part of these challenges pertains directly to data and can be addressed at L1 thus increasing flexibility and productivity due to the implementation at the shortest distance to the signal initiators [13].

One possible solution is *the creation of exemplary FBs* (or reusable software components in general). Such standardized components

would encapsulate best practices and methodologies for maintaining data quality, granting users greater autonomy and flexibility to customize data quality measures to their needs. This approach might empower organizations to implement effective data quality strategies and promotes a more standardized and customizable framework for managing and improving data quality in diverse contexts. Due to the computational limitations, L1 is often not considered as a solution for higher complexity data issue problems. Nonetheless, there are successful attempts to implement complex mathematical models at L1 [13]. The question is how already existing ready-to-use basic building blocks can be applied for it.

Furthermore, *data flow visualization* at L1 can be beneficial for the identification of data issues. By implementing intuitive and informative visualizations, operators and analysts can gain better insights into the data flow and identify potential issues despite the high data volumes including data anomalies and bottlenecks.

Efficient metadata is another challenge to be addressed not only with standardization but also by increasing adaptability and responsiveness to current needs. For instance, when dealing with metadata for a temperature sensor, essential information may encompass date and time of the measurement, the temperature unit used (e.g., degree Celsius or degree Fahrenheit), details regarding the sensor's calibration, and specific equipment or location data. However, *the metadata framework must be flexible and subject to change*, based on the insights and expertise of domain-specific experts. By embracing a dynamic approach to metadata management, industries can effectively tailor their metadata requirements to suit unique contexts and evolving technologies, empowering them to make data-driven decisions that lead to improved operational efficiency and enhanced overall performance.

Addressing data quality issues at the L1 offers a significant advantage in terms of speed of reaction and problem-solving in industrial automation. L1's direct access to sensor data allows for swift identification and troubleshooting, minimizing downtime and disruptions. In contrast, the higher-level L2 focuses on optimization and may not have real-time visibility into sensor values, potentially leading to delays in issue detection. Prioritizing data quality at L1 ensures reliable information is passed to L2, facilitating better decision-making and improving overall system reliability and efficiency. The optimal solution would be data quality measures integration at both levels resulting in a comprehensive framework for managing and optimizing data quality in industrial automation.

Data quality requirements indeed exhibit variations across different industrial domains, with certain elements assuming greater importance in specific sectors. Although various industries have developed effective practices, the real challenge lies in collecting and sharing this knowledge. The reluctance to disclose proprietary information creates barriers to collaboration and impedes the establishment of comprehensive best practices and metadata frameworks. To progress in this area, fostering cross-industry learning and cooperation is vital, enabling the development of standardized metadata frameworks that can benefit all. However, the willingness of developers to participate in knowledge-sharing practices remains a key obstacle that must be addressed to achieve a more unified approach to metadata management across diverse industrial domains.

## 5 CONCLUSION

In this paper, we describe the issues observed in practice in a plant-building company that can be addressed at L1 including data compression, metadata formation, and knowledge transfer issues. Based on the identified areas, future work should focus on several areas. Firstly, research efforts can be directed toward the development and implementation of FBs focusing on data quality validation and lossless data compression. Secondly, the creation of standardized metadata frameworks that cater to diverse industrial domains and facilitate seamless data sharing and interoperability remains a critical problem. Additionally, promoting knowledge sharing and collaboration among stakeholders can help overcome the challenges related to knowledge transfer, encouraging customers to contribute valuable insights and feedback for continuous improvement.

## ACKNOWLEDGEMENTS

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged. Furthermore, this work was supported by the Austrian Research Promotion Agency (FFG) in the frame of the project ConTest [888127]. Finally, we also want to thank our interview partners for their time and valuable input.

## REFERENCES

[1] 2003. IEC 62264: Enterprise-control system integration. International Electrotechnical Commission. https://www.iso.org/standard/31332.html Standard.
[2] Ahmed Abdulhasan Alwan, Mihaela Anca Ciupala, Allan J Brimicombe, Seyed Ali Ghorashi, Andres Baravalle, and Paolo Falcarin. 2022. Data quality challenges in large-scale cyber-physical systems: A systematic review. *Information Systems* 105 (2022), 101951.
[3] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–52.
[4] Harald Foidl and Michael Felderer. 2023. An approach for assessing industrial IoT data sources to determine their data trustworthiness. *Internet of Things* 22 (2023), 100735.
[5] IEC. 2013. IEC 61131 - Programmable controllers, Part 3: Programming languages: Edition 3.0. www.iec.ch
[6] IEC TC65/WG6. 2012. IEC 61499-1, Function Blocks - part 1: Architecture: Edition 2.0. www.iec.ch
[7] Hyo-Sang Lim, Yang-Sae Moon, and Elisa Bertino. 2010. Provenance-based trustworthiness assessment in sensor networks. In *Proceedings of the Seventh International Workshop on Data Management for Sensor Networks*. 2–7.
[8] Guolin Lyu and Robert William Brennan. 2021. Towards IEC 61499-Based Distributed Intelligent Automation: A Literature Review. *IEEE Transactions on Industrial Informatics* 17, 4 (2021), 2295–2306. https://doi.org/10.1109/TII.2020.3016990
[9] László Monostori. 2014. Cyber-physical production systems: Roots, expectations and R&D challenges. *Procedia CIRP* 17 (2014), 9–13. https://doi.org/10.1016/j.procir.2014.01.001
[10] Ricardo Perez-Castillo, Ana G Carretero, Moises Rodriguez, Ismael Caballero, Mario Piattini, Alejandro Mate, Sunho Kim, and Dongwoo Lee. 2018. Data quality best practices in IoT environments. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*. IEEE, 272–275.
[11] Ivan Miguel Pires, Nuno M Garcia, Nuno Pombo, Francisco Flórez-Revuelta, Natalia Díaz Rodríguez, et al. 2016. Validation techniques for sensor data in mobile health applications. *Journal of Sensors* 2016 (2016).
[12] Sagar Sen, Erik Johannes Husom, Arda Goknil, Simeon Tverdal, Phu Nguyen, and Iker Mancisidor. 2022. Taming data quality in AI-enabled industrial internet of things. *IEEE Software* 39, 6 (2022), 35–42.
[13] Thomas Trenner, Jörg Neidig, Rolf Findeisen, and Stefan Streif. 2014. Einsatz cyber-physischer Systeme im Echtzeitkontext. *Automation* (2014), 311–324.