IoT Data Quality Assurance Framework for Reliable Process Analytics

Andrea Maldonado[1,*], Adrian Bazan-Muñoz[2], Lukas Malburg[3,4] and Yannis Bertrand[5]

[1]University of Bayreuth, Bayreuth, Germany
[2]UCASE Software Engineering Group, University of Cádiz, Puerto Real, Cádiz, Spain
[3]Artificial Intelligence and Intelligent Information Systems, Trier University, Germany
[4]German Research Center for Artificial Intelligence (DFKI), Germany
[5]Research Group Business Informatics, Hasselt University, Hasselt, Belgium

Abstract

Data quality is a major concern in all data analysis tasks. In IoT-enhanced business processes, data quality issues often arise from faulty sensor measurements, which are propagated through event abstraction to incorrect process decisions. In this paper, we investigate the main challenges in data quality management for IoT-enhanced business processes and isolate four subchallenges. Each subchallenge is first defined and put in perspective with relevant literature and representative real-world use cases before we present five overarching opportunities to address the subchallenges. Finally, we conclude the paper with a research agenda, making recommendations over multiple time horizons to improve data quality management in IoT-enhanced business processes.

Keywords

IoT-enhanced Business Processes, Data Quality Management, Internet-of-Things, Expert-in-the-Loop

1. Introduction

Data quality is a concern in all data analysis tasks. Data quality issues (DQIs) typically arise from design flaws in information systems (e.g., mismatches between the designed system and user requirements) or from logging issues (e.g., connection problems that introduce noise in sensor data). Previous research identified various dimensions to describe and quantify data quality, such as: accuracy, timeliness, precision, completeness, reliability, and error recovery [1]. In particular, [2] makes a distinction between intrinsic data quality, which refers to the *correctness* of the data (e.g., data accuracy) and other aspects of data quality, which can be described as the *convenience for use* of the data (e.g., data timeliness, granularity) [2, 3, 1]. Correctness refers to the extent to which the data represent reality. In this sense, incorrect data do not accurately represent what actually happened, e.g., data are missing or have incorrect values. This is the case when, for instance, the activity label logged does not correspond to the activity that was executed. Convenience for use refers to the extent to which the data are readily usable for a given (analysis) task. This is the case when, for instance, a sensor outputs one or multiple measurements per second, generating huge quantities of data. Note that data can be correct but difficult to analyse, e.g., due to an ill-adapted granularity level or a high volume.

In this paper, we focus on the first, stricter notion of DQIs, i.e., DQIs as incorrect data. This is because we assume that the fact that data are at a low granularity level and that huge amounts of data are collected is addressed by analysis techniques. For instance, event abstraction (e.g., CEP) can be used to derive higher-level events from raw sensor output, which solves the issues of granularity and data quantity.

The presence of DQIs makes data analysis more difficult because it entails more data preprocessing to obtain sufficient results. Without proper data preprocessing, DQIs have adverse effects on the output

BP-meet-IoT workshop 2026
*Corresponding author.
andrea.maldonado@uni-bayreuth.de (A. Maldonado); adrian.bazan@uca.es (A. Bazan-Muñoz); lukas.malburg@uni-trier.de (L. Malburg); yannis.bertrand@uhasselt.be (Y. Bertrand)
0009-0009-8978-502X (A. Maldonado); 0000-0001-5182-4905 (A. Bazan-Muñoz); 0000-0002-6866-0799 (L. Malburg); 0000-0002-6407-7221 (Y. Bertrand)

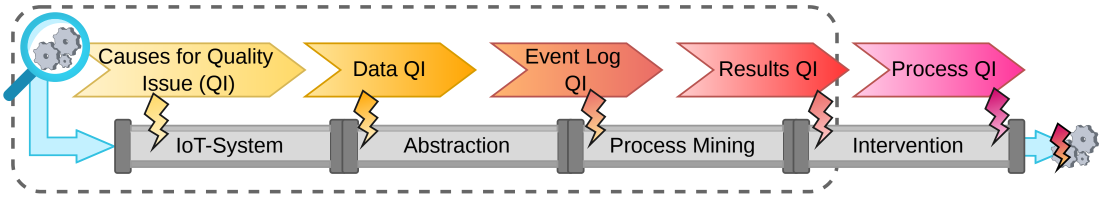

Figure 1: DQIs propagate along the steps of the process analysis pipeline.

of data analysis tasks. Examples include, e.g., training an ML model with wrong input will lead to a wrong model; more data is necessary to help the model separate the actual signal from noise; the process model is more complex due to noise. The reach of the adverse effects of DQIs does not stop with poor data analysis results and has negative consequences on subsequent business decisions [4, 5]. For instance, wrong predictions from an AI method will lead to incorrect decisions, such as suboptimal resource allocation; collecting more data can be time- and resource-consuming; and it is difficult for process owners to trust process mining results.

More specifically, we view the impact of IoT DQIs in BPM as a propagation in the analytics pipeline, as shown in Figure 1, where DQIs originate at the hardware level and propagate through subsequent analysis stages [6]:

- In IoT systems, DQIs are often due to cheap hardware in difficult environments. These faults stem from poor sensor calibration, network issues, low battery, temperature/humidity variations affecting the working of sensors, etc. [1, 7]. In addition to this, software errors can also tamper correctly registered readings, such as incorrect units, Such sensor faults cause IoT *Data QI*s, such as missing IoT events, noise, incorrect timestamps, etc.
- When IoT events are abstracted to derive process events or contextualise an event log, existing IoT DQIs lead to *Event Log QI*s, like incorrect case IDs, incorrect activity labels, or missing events.
- Processing these faulty higher-level process data will very likely yield *Results QI*s. That is, incorrect conclusions about the process and how it is being executed, including complex process models, wrong activity orders, incorrect predictions about ongoing cases, etc.
- Finally, these poor analyses will mislead process owners and cause them to make wrong decisions about how to redesign the process or steer specific process cases, resulting in *Process QI*s instead of improvement.

To address this issue, we split the major challenge of data quality management in IoT-enhanced processes into four sub-challenges: 1) *detecting and repairing IoT DQIs*; 2) *performing root-cause analysis for identifying systemic sources for DQIs*; 3) *enabling real-time monitoring*; and 4) *strengthening robustness and trust in process analytics and automation pipelines.*

In the remainder of the paper, we first outline each subchallenge in more detail in Section 2. Then, in Section 3, we discuss literature relevant to each of the subchallenges and illustrate how these subchallenges affect two real-life use cases. Thereafter, section 4 presents overarching opportunities to tackle the subchallenges. Based on the opportunities identified in the previous section, Section 5 lists concrete recommendations to improve data quality in IoT-enhanced process environments. Finally, Section 6 summarises our findings and concludes the paper.

2. Subchallenges

Ensuring data quality in IoT-driven process analytics involves multiple interconnected challenges. To structure this problem, a set of key subchallenges is identified, spanning the process analysis pipeline from data acquisition and quality assessment to real-time monitoring and its impact on downstream analytics. Together, they provide a focused view of the main research directions required to achieve reliable IoT-based process analytics.

2.1. Detecting and repairing IoT DQIs

The first and principal challenge in IoT data quality is to be able to detect DQIs and subsequently to repair or, at least, mitigate them. Detecting IoT DQIs primarily entails distinguishing data points that correctly represent reality from data points that don't. The difficulty of a detection task can range from trivial, e.g., when some data points are missing, to extremely difficult, e.g., when incorrect data do not differ from correct data. Expert knowledge is often required to define and recognise more subtle DQIs. Once detected, DQIs should be handled. There are three main ways to address DQIs: 1) ignoring DQIs, which can be acceptable if it is more costly to treat them than to leave them in the data; 2) removing DQIs, which is often done when no obvious repair approach is known; 3) repairing DQIs, which consists of trying to identify the actual value of an incorrect data point and replace the wrong value with the correct one. This third option is usually the best, since it allows the user of the data to use the whole dataset (not the case in approach 2) without DQIs (not the case in approach 1), but it is also the most complex, since it requires being able to infer, from an incorrect data point, what the actual value of that data point should have been. This usually relies on either domain knowledge or the analysis of data points with a similar context, or both.

2.2. Identifying root-causes in systemic DQIs

Identifying the root-cause of DQIs is a pivotal research challenge because DQIs will persist unless their underlying systemic source is permanently resolved. In the simplest case, systemic sources of DQIs include improperly calibrated sensors or faulty and/or damaged actuators. However, a systemic source of a DQI may also arise from a combination of multiple faulty sensors and actuators and may depend on the currently executed process. Consequently, root-cause analysis typically relies on domain expert knowledge to identify the root-cause in large, highly interconnected systems, such as IoT environments [8]. In addition, the high volume, variety, and velocity of IoT sensor data, which must be processed and accurately labeled for AI-driven techniques, raise further concerns regarding real-time readiness, preprocessing and training effort, and, ultimately, the suitability of the resulting models. Addressing these shortcomings typically requires hybrid AI systems that enable informed, data-driven root-cause analysis. Such systems combine purely data-driven AI techniques with explicitly modeled semantic knowledge, such as expert knowledge about typical failures and errors, their root-causes, and potential correlations and dependencies within IoT sensor data.

2.3. Enabling real-time stream monitoring

The challenge of enabling real-time monitoring lies in detecting and mitigating DQIs during data ingestion, ensuring that only reliable information is propagated to downstream information processing systems. Unlike batch processing approaches, real-time monitoring must operate under the constraints of high-velocity and non-stationary IoT data streams, where contextual conditions may change abruptly or gradually, potentially invalidating static quality control mechanisms [9]. This challenge is further exacerbated by the inherent asynchrony of distributed sensor networks, which frequently leads to out-of-order event arrivals that must be reconciled without introducing significant processing latency, as delays directly undermine the benefits of real-time analytics. Consequently, a key research objective is the design of monitoring architectures capable of maintaining end-to-end observability across the Edge-Cloud continuum [10], providing immediate and continuous feedback on the quality and reliability of incoming data to prevent the propagation of erroneous information to higher-level process models and analytics pipelines [11].

2.4. Strengthening robustness, and trustworthiness in automation pipelines

Furthermore, variations in data, for example, due to DQIs, often heavily affect the results of process analysis pipelines [12]. The effects of varying data characteristics, including complexity, incompleteness, and data irregularities, typically challenge the robustness of algorithms and thus the reliability of their results. This introduces representation bias, as analysis results may not represent the underlying characteristics of the data distribution. Particularly in IoT systems, process steps are carefully extracted from sensor data to form activities for further analysis. As a result, DQI may introduce uncertainty and noise into activity results, further undermining the robustness of automated process analysis pipelines. Additionally, although automation pipelines facilitate feasibility and real-time analysis, robustness problems are further aggravated by the aggregation of biased datasets and results. As process analytics algorithms are trained on specific datasets, these datasets need to representatively cover the high volume, variety, and velocity of real IoT scenarios to enable automation pipelines in production to scale under pressure.

3. Current Landscapes in IoT Process Analytics

This section reviews the current state of the art on IoT data quality in the context of process analytics. The analysis is structured along two complementary perspectives, namely a research perspective, which examines existing methods and approaches proposed in the literature, and a practical perspective, which highlights real-world applications and challenges observed in operational settings.

3.1. Research Perspective: Methods and Frameworks

From a research perspective, existing work addresses IoT data quality across several key dimensions, including detection and repair of DQIs, root-cause analysis, real-time monitoring, and the robustness of process analytics pipelines. The following subsections provide an overview of the main approaches and trends within each of these areas.

3.1.1. Detection and Repair Techniques

IoT DQI detection and repair is a heavily researched field. Multiple recent reviews summarise the state of the art [13, 14]. In [13], the authors review the state of the art in IoT data quality, viewing data quality as the extent to which the data correctly represent reality. In their paper, they report the most frequent types of DQIs encountered in the broader IoT literature: Outliers, Missing data, Bias, Drift, Noise, Constant value, Uncertainty and Stuck-at-zero. Later in [13], the authors review DQI detection techniques, reporting 15 different approaches, mostly aiming at identifying outliers. Finally, this paper also report seven distinct DQI repair techniques, addressing either outliers or missing data. So while a plethora of techniques exist, multiple challenges remain, given that 1) most techniques only address one DQI; and 2) some DQIs are not tackled by any detection or repair technique. More recently, [14] take a broader perspective at data quality, surveying data preprocessing techniques in general. This includes approaches to prepare data for machine learning, such as normalisation or feature selection. They report 57 data preprocessing techniques and benchmark them for a time series forecasting problem using LSTM neural networks.

A more focused look at DQIs in the IoT BPM literature in [6] yielded a slightly different picture, with noise being the most often reported DQI type, and missing data and outliers the second and third most frequent, respectively. This discrepancy between the general IoT literature and the IoT BPM field could be due to a lower awareness of the different types of DQI, leading to various DQI types being “labelled” as noise. Regarding repair techniques, [15] propose an expert-driven approach to correct DQIs in IoT event logs.

3.1.2. Hybrid and Data-Driven Root-Cause Analysis

Much research has been conducted in the area of root-cause analysis in IoT environments and especially in the context of production. Stertz et al. [16] present an approach to trace the root-cause of concept drifts by correlating them with sensor data. The approach demonstrates how sensor event sequences can be aligned with process events to explain process deviations. A first approach addressing the challenge of using hybrid AI systems combining data-driven AI techniques with expert knowledge has been presented by Steenwinckel et al. [17]. In this approach, a knowledge-driven root-cause analysis approach in IoT is presented, which can explain anomalies based on an experience-based approach combined with symbolic-driven neural reasoning and formalized expert knowledge. Klein et al. [8] present a similar approach in which a knowledge graph is used to detect anomalies and to determine the root-cause of them. Ghalibafan et al. [18] propose the eRooMiner approach, which is a data-driven approach combining machine learning and AI techniques to determine root-causes based on DQIs in event logs. A quite similar approach has been presented by Andrews et al. [19], which is based on the Odigos Root-Cause reference framework. The goal of the approach is to deal with DQIs on a prognostic and diagnostic level.

3.1.3. Adaptive Real-Time Monitoring Architectures

Research on real-time IoT data quality monitoring has evolved from static analysis approaches toward distributed and adaptive frameworks capable of operating under streaming conditions [20, 21]. Early work established the technical foundations for integrating process-aware systems with live sensor streams [11], which were later reflected in domain-specific systems such as BPSmart-Care [22]. Subsequent approaches focused on near real-time detection and remediation of DQIs using techniques such as Case-Based Reasoning (CBR) [23]. From an architectural perspective, frameworks such as ContinuumConductor optimize the placement of data cleaning and process discovery tasks across the Edge-Cloud continuum to support low-latency processing [10, 24]. Other research directions address the detection of abrupt and gradual changes in sensor streams through drift detection methods such as McDiarmid Drift Detection Method (MDDM) [9], as well as the uncertainty and asynchrony of IoT data using probabilistic and sliding-window techniques for handling incomplete and out-of-order events [25, 26]. More recently, systems such as IoT Miner [27] explore the integration of large language models (LLMs) and unsupervised learning techniques to automatically transform raw sensor data into high-level event logs suitable for online process analytics.

3.1.4. Algorithmic Robustness and Trustworthiness

To investigate robustness, reliability, and trustworthiness of process analytics algorithms in offline scenarios, several works recur to understanding the underlying distribution of input data to create controlled datasets to use in training and testing of process analysis algorithms. Regarding reliability and trustworthiness, the list concerns current challenges and opportunities for fault and repair methods, particularly on data availability [28, 29, 30]. Different domains exhibit different fault frequencies. On the one hand, Stevens et al. [31] assessed the vulnerabilities in current outcome prediction models. To improve the realism of their approach, Stevens et al. [32] presented a method that employs manifold learning the underlying distribution of a numerical or categorical attribute value to generate out-of-distribution samples during adversarial attacks. On the other hand, Maldonado et al. [33] work focuses on the control-flow perspective, generating data with intentional structural event characteristics for training and testing of process discovery algorithms. They follow up by presenting an approach to explain the robustness of process discovery algorithms against variations in these structural event characteristics. [34] Although these works propose methods for testing and training robust methods, none are specialized for IoT real-time scenarios. To this end, Know Your Streams [12] was tuned to generate event data for online scenarios, controlling the level of e.g. out-of-order events, concurrent activities, incomplete cases, and concept drifts, which can all be caused by DQI. Furthermore, AVOCADO [35] proposes a framework for evaluating method robustness using controlled online datasets. Still, the link between DQI and data characteristics, as well as their relation to robustness, reliability, and trustworthiness, remains to be investigated, particularly in IoT scenarios. Another remaining gap is to integrate expert domain knowledge and differentiation of domains (E.g. manufacturing vs. aircrafts and nuclear plants) products [28].

3.2. Practical Perspective: Case Study Evaluations

From a practical perspective, IoT data quality challenges manifest in real-world systems where heterogeneous data sources, complex environments, and operational constraints introduce additional sources of variability and uncertainty. The following case studies illustrate how these challenges arise in practice, how they are addressed, and which requirements the quality management approaches need to satisfy. Table 1 summarizes the main characteristics, requirements, DQIs, mitigation strategies, and research opportunities identified across both cases.

3.2.1. Case study 1: Pervasive healthcare system in a nursing home for proactive prevention of COVID-19

This case study considers a pervasive healthcare system deployed in a nursing home for the proactive prevention of COVID-19 [36]. The system integrates heterogeneous IoT data sources, including wearable devices providing physiological measurements, environmental sensors monitoring air quality conditions, and contextual data describing proximity and occupancy. These data are combined within a data fabric and digital twin infrastructure supporting real-time analytics, inference generation, and automated actions such as alerts and notifications.

The objective is to ensure reliable digital twin representations and decision-support outputs through continuous data quality assessment. Several IoT DQIs are observed, including missing values, sensor noise, inaccurate physiological or environmental measurements, and errors in proximity detection caused by imprecise location or occupancy data. In addition, inconsistencies introduced during data fusion may propagate through the pipeline and affect downstream inferences.

These issues may lead to false positives or false negatives in infection-risk detection, unreliable environmental assessments, and reduced trust in automated recommendations. From a process analytics perspective, incorrect inferred events negatively affect both preventive process execution and analysis.

The case study addresses several challenges identified in Section 2. Automated quality assessment mechanisms are applied throughout the pipeline to detect missing values, out-of-range measurements, and inconsistencies across sensors. Domain knowledge is leveraged to identify semantically implausible situations and support repair strategies such as data imputation, smoothing noisy measurements, and correcting inconsistent values using cross-sensor correlations.

Furthermore, the system continuously monitors IoT data quality in real time, particularly after data fusion stages, by computing metrics such as accuracy and completeness based on ISO/IEC 25024. This monitoring helps prevent the propagation of low-quality data and improves the robustness and trustworthiness of process analytics pipelines. However, limitations remain regarding the limited set of quality dimensions considered, short evaluation periods, and the handling of dynamic context changes.

3.2.2. Case study 2: Chemical manufacturing

This case study considers a chemical manufacturing process in which raw materials are mixed and filtered in a sensor-monitored tank. The process is supported by a Manufacturing Execution System (MES) logging process events at different granularity levels, while IoT sensors capture physical parameters such as pressure and flow measurements.

The objective of the analysis is to perform process mining in order to identify sources of variation and potential inefficiencies in the manufacturing process. Unlike the previous case study, the analysis is performed offline and does not require real-time processing.

Several DQIs are observed in both MES and sensor data. In the MES, these include incorrect or inconsistent activity labels, imprecise activity granularity, timestamp ambiguities caused by differences between execution and reporting times, and inconsistencies related to daylight-saving time changes. Sensor data also exhibit quality problems, particularly noise and lossy compression mechanisms that omit repeated constant values.

These issues negatively affect process analytics results. Timestamp inconsistencies lead to incorrect correlations between IoT events and process activities, while incorrect labels and granularity problems reduce the reliability of process discovery results. In addition, the inability to correctly identify anomalous production patterns may ultimately affect batch quality.

This case study mainly illustrates the challenges of detecting, repairing, and analysing the root causes of IoT DQIs. Detection primarily relies on visual data exploration and iterative discussions with domain experts. Repair strategies include timestamp harmonization, label standardization, forward-fill imputation of missing sensor values, and denoising techniques based on rolling averages. Furthermore, root-cause analysis is supported through expert knowledge to identify systemic sources of recurring data quality problems.

Table 1
Compact comparison of IoT data quality challenges, mitigation strategies, and research opportunities in the case studies.

Dimension	Case Study 1: Pervasive Healthcare	Case Study 2: Chemical Manufacturing
Context & goal	Nursing-home digital twin and data fabric for real-time COVID-19 risk prevention and alerts.	Sensor-monitored tank and MES logs for offline process mining and variation analysis.
Key requirements	Real-time quality monitoring, reliable and timely alerts, privacy-preserving data handling, and scalability to heterogeneous sensor streams.	Reliable event–activity correlation, robust process analysis, explainable results, and consistent event-log representation.
Observed DQIs	Missing values, sensor noise, imprecise proximity/location data, and propagated fusion errors.	Timestamp ambiguity, daylight-saving shifts, and lossy compression.
Process impact	DQI can reduce digital-twin fidelity and undermine inferred alerts.	Timestamp shifts and compression weaken event–activity correlation.
Detection	Mostly rule-/threshold-based quality checks with limited context adaptation.	Visual exploration and expert consultation.
Repair	Domain imputation, smoothing, and cross-sensor consistency repair.	Timestamp harmonization, label standardization, and rolling-average denoising.
Expert role	Continuous validation and repair prioritization in feedback loops.	One-shot expert-driven root-cause analysis.
Research opportunity	**Online adaptivity:** proactive, context-aware DQI management in dynamic IoT settings.	**Root-cause analysis:** hybrid expert- and data-driven identification of systemic DQI sources.

4. Gaps and opportunities

Despite the significant progress outlined in the previous section, several limitations remain in current approaches to IoT data quality management for process analytics. These limitations reveal important research gaps and opportunities that must be addressed to enable reliable and scalable IoT-driven process analytics. The following subsections discuss these gaps in relation to the previously identified subchallenges. These opportunities are positioned in Figure 2, composed by the process analytics pipeline, as presented in Section 1, and the process quality management lifecycle [37].

4.1. Opportunity 1 – Reliable Process Representation

As presented in Section 3.1.4, so far, the state of the art lacks a systematic investigation, aligning raw DQIs, such as sensor noise, in event log *abstraction* with higher-level metrics of reliability, robustness of the *process mining results*. While tailored validated solutions for handling DQIs of specific processes exist, the field lacks data quality assessment approaches to quantify the statistical significance and certainty of data representing a corresponding process. Representativeness and coverage of IoT data in respect to the process, avoids overgeneralization of process analytics approaches, since one size does not fit them all. For example, nuclear plants processes might require stricter robustness and safety-critical reliability in analysis results compared to standard manufacturing. This way, our process analytics approaches shall be capable of processing event streams in a contextually reliable and robust manner for high-stakes decision making. Thus, there is a significant opportunity to develop robustness-assessing preprocessing frameworks that *detect*, *quantify* and *repair* sampling vulnerabilities in DQIs.

4.2. Opportunity 2 – Generalization of Approaches

Many techniques were proposed to detect and repair IoT DQIs, but most of them are limited to a subset of DQIs [13]. In particular, detection techniques mainly focus on outlier detection, while repair techniques focus on noise and missing data. This makes it difficult to accurately *detect and repair* DQIs in IoT environments, where multiple DQI types typically co-exist.

Hence, there is a need for more general-purpose approaches. These approaches could either consist of techniques that directly detect or repair various types of DQIs, or of more elaborate frameworks combining multiple DQI detection or repair techniques. In the first case, a possibility is to use refined anomaly detection techniques to detect data points that differ from the rest of the data and isolate them as likely containing DQIs. The second path is to develop a broader framework combining different DQI detection or repair techniques to make it possible to detect and repair a wide range of DQI types.

4.3. Opportunity 3 – Root Cause Analysis

Future work offers substantial opportunities to advance root-cause analysis for DQIs by going beyond purely data-driven AI approaches. While current data-driven neural AI methods only rarely make systematic use of expert knowledge to identify the underlying sources of DQIs, such knowledge represents a valuable resource for improving both the accuracy and interpretability of root-cause identification. By combining the strengths of data-driven AI techniques with domain knowledge, hybrid neuro-symbolic AI approaches can be developed that are better suited to diagnose systemic causes of data quality problems. This integration promises not only more reliable, trustworthy, interpretable, and explainable diagnoses [8] but also more actionable insights for organizations seeking to improve their data management processes.

Beyond identifying the root causes of DQIs, there is also a significant opportunity to investigate their actual impact. Although it is commonly assumed that poor data quality negatively affects analysis results, following the principle of "garbage in, garbage out," the extent and nature of this impact often remain insufficiently understood. Different types of DQIs may influence analytical outcomes in different ways, and not all issues may justify the same level of repair effort. A more profound understanding of these effects would make it possible to assess when correcting DQIs is beneficial and when the costs of repair may outweigh the potential improvements. This perspective can support more informed,

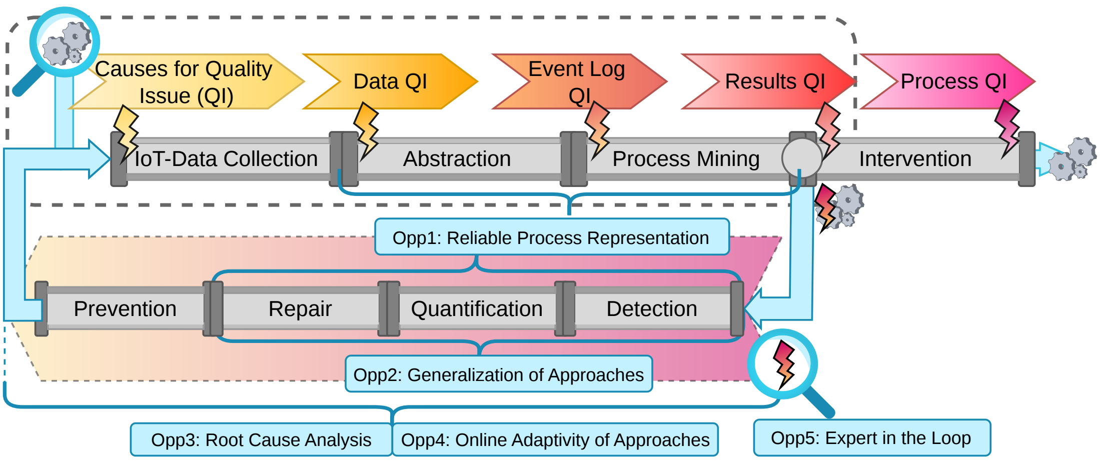

Figure 2: Relationships between five identified opportunities, the process analytics pipeline and process data quality management lifecycle [37].

cost-effective, and targeted *data quality management* decisions.

4.4. Opportunity 4 – Online Adaptivity of Approaches

Current advances in real-time IoT data quality monitoring create significant opportunities for more adaptive and context-aware process analytics pipelines. Existing approaches have demonstrated the feasibility of monitoring selected data quality dimensions in streaming environments, but future research can extend these capabilities toward more comprehensive quality assessment mechanisms.

A key opportunity lies in developing adaptive monitoring frameworks capable of dynamically adjusting quality assessment strategies according to evolving environmental and process conditions. Rather than relying on static rules or predefined thresholds, next-generation approaches should support context-aware quality management in highly dynamic IoT environments.

Another promising direction is the integration of data quality monitoring with downstream process analytics, enabling systems to assess not only the presence of DQIs, but also their impact on process analysis reliability and decision-making outcomes.

Furthermore, there is strong potential for proactive and predictive *data quality management* approaches capable of anticipating quality degradation before it propagates through the system. In parallel, future research should move toward end-to-end monitoring frameworks spanning the complete IoT data lifecycle, from edge-level acquisition to cloud-based processing and analysis.

4.5. Opportunity 5 – Expert in the Loop

A critical challenge impacting all parts of the *process analytics pipeline*, as well as the *data quality management lifecycle*, is the integration of expert domain knowledge in the *feedback loop*. Purely automated approaches struggle with DQI subtlety. DQI detection requires a fine, specialized calibration due to the high interconnection between the propagation of DQIs and process quality issues in IoT environments. Approaches that treat data characteristics as purely statistical phenomena struggle to fulfil requirements, which vary substantially across use-cases and domains. Therefore, a clear opportunity remains to create hybrid process data quality management systems. That is, frameworks combining statistical soundness, as well as interactive and continuous integration of expert domain knowledge. By involving the expert-in-the-loop, researchers can develop more contextually aware repair mechanisms that reflect domain-specific context to infer correct values in specialized industrial sectors. This shift ensures that representation bias in current analytics methods, mentioned in Section 3.1.4, is mitigated by a nuanced high-level understanding that only human domain experts possess.

5. Proposed interventions and recommendations

Based on the identified gaps and challenges, this section outlines a set of research and practical recommendations to advance IoT data quality management for process analytics. These recommendations are structured along short-, mid-, and long-term perspectives, reflecting different levels of maturity and implementation effort.

5.1. Short-term recommendation: Validation

In the short term, efforts should focus on addressing immediate limitations in current approaches. First, there is a need to develop detection and repair techniques for DQIs that are currently underexplored, beyond commonly addressed cases such as noise and missing data. Additionally, data quality validation mechanisms should be introduced as early as possible in the process analysis pipeline, particularly at the level of sensors and devices, to prevent the propagation of erroneous data.

Furthermore, existing monitoring systems should be extended to incorporate a broader set of data quality metrics, enabling a more comprehensive assessment of data reliability. Finally, initial steps should be taken toward systematically linking DQIs with data characteristics and evaluating quality control mechanisms within process analysis pipelines.

5.2. Mid-term recommendation: Adaptation

In the mid term, research should move toward more integrated and generalizable solutions. This includes the development of general-purpose approaches for detecting and repairing DQIs across multiple types of DQIs. In parallel, process domain knowledge should be systematically captured and incorporated into root-cause analysis methods to improve their effectiveness.

From a system perspective, there is a need to design architectures that support data quality assessment across multiple stages of the data pipeline, enabling the identification of where and how quality degradation occurs. Such architectures should also support dynamic adaptation of data quality rules in response to changing context conditions. Moreover, stronger links should be established between DQIs and their impact on process analytics outcomes, while ensuring the integration of domain expertise and differentiation across application domains.

5.3. Long-term recommendation: Prevention

In the long term, the focus should shift toward proactive and predictive data quality management strategies, preventing issues before they occur. These include developing end-to-end monitoring and correction mechanisms that control the propagation of DQIs throughout the entire system.

Future research should also explore the use of digital twins of data infrastructures to model and monitor data quality dynamically, enabling continuous assessment and correction. In addition, evaluating the real-time impact of data quality degradation on process analytics remains an open challenge. Finally, the integration of domain expertise and the differentiation between application domains should be further strengthened to support more robust and context-aware solutions.

6. Final synthesis and conclusions

In this work, we have evaluated current challenges regarding IoT DQIs and how these affect the reliability of process process analysis pipelines. As we explored the propagation of DQIs at different steps, we split this major challenge into four subchallenges, spanning over the entire process analytics pipeline and data quality management lifecycle, from data acquisition, quality assessment, to analysis. We examined current methods and approaches to address these challenges, and identified gaps in the state of the art, while also illustrating two case studies for practical perspectives on possible solutions and opportunities. From our analysis of the field, we identified five major opportunities for future research, which we summarize in the following: First, improving *reliable process representation* in the presence of DQIs, including missing and noisy data. Particularly, because uncertainty and complexity are often present during data abstraction of real-world processes. Second, the *generalization of approaches* to ensure that methods can be applied across different domains and contexts. Third, to identify the underlying causes of DQIs, opportunities in *root-cause analysis* arise. Fourth, developing *online adaptive approaches* that adapt to changing conditions in real-time opens multiple venues through the data quality management lifecycle. Overall, we want to highlight the importance of involving human *experts in the loop*. By leveraging their domain knowledge, the reliability of analysis results as well as their interpretations become more useful and reliable. With these opportunities in mind, we provide a short-term, mid-term, and long-term roadmap for researchers and practitioners. In the short-term, the focus should be on implementing a comprehensive DQI detection and thorough repair mechanisms. Moving into the mid-term, research should aim at integrative adaptive architectures tracking quality in a human conscious manner across the entire pipeline. In the long-term, envision proactive, self-healing systems using digital twins and end-to-end correction systems for real-time transparency and trust in results and process quality.

References

[1] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, Data quality in internet of things: A state-of-the-art survey, Journal of Network and Computer Applications 73 (2016) 57–81.

[2] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, JMIS 12 (1996) 5–33.

[3] C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques, DCSA, Springer, 2006.

[4] A. Haug, F. Zachariassen, D. Van Liempd, The costs of poor data quality, Journal of Industrial Engineering and Management (JIEM) 4 (2011) 168–193.

[5] M. Eppler, M. Helfert, A classification and analysis of data quality costs, in: International Conference on Information Quality, MIT Cambridge, 2004, pp. 311–325.

[6] Y. Bertrand, R. Van Belle, J. De Weerdt, E. Serral, Defining data quality issues in process mining with iot data, in: International Conference on Process Mining, Springer, 2022, pp. 422–434.

[7] A. Goknil, P. Nguyen, S. Sen, D. Politaki, H. Niavis, K. J. Pedersen, A. Suyuthi, A. Anand, A. Ziegenbein, A Systematic Review of Data Quality in CPS and IoT for Industry 4.0, ACM Computing Surveys 55 (2023) 1–38.

[8] P. Klein, L. Malburg, R. Bergmann, Combining informed data-driven anomaly detection with knowledge graphs for root cause analysis in predictive maintenance, Engineering Applications of Artificial Intelligence 145 (2025) 110152. doi:10.1016/j.engappai.2025.110152.

[9] A. Pesaranghader, H. L. Viktor, E. Paquet, Mcdiarmid drift detection methods for evolving data streams, in: 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–9. doi:10.1109/IJCNN.2018.8489260.

[10] H. Reiter, J. Edinger, M. Kabierski, A. Koschmider, O. Landsiedel, A. Lepsien, X. Lu, A. Marrella, E. Serral, S. Schulte, F. Tschorsch, M. Weidlich, W. Hasselbring, Continuumconductor : Decentralized process mining on the edge-cloud continuum, ArXiv abs/2512.07280 (2025). URL: https://api.semanticscholar.org/CorpusID:281528724.

[11] C. Janiesch, A. Koschmider, M. Mecella, B. Weber, A. Burattin, C. Di Ciccio, G. Fortino, A. Gal, U. Kannengiesser, F. Leotta, F. Mannhardt, A. Marrella, J. Mendling, A. Oberweis, M. Reichert, S. Rinderle-Ma, E. Serral, W. Song, J. Su, V. Torres, M. Weidlich, M. Weske, L. Zhang, The internet of things meets business process management: A manifesto, IEEE Systems, Man, and Cybernetics Magazine 6 (2020) 34–44. doi:10.1109/MSMC.2020.3003135.

[12] A. Maldonado, C. Imenkamp, H. Reiter, T. Seidl, W. Hasselbring, M. Werner, A. Koschmider, Know your streams: On the conceptualization, characterization, and generation of intentional event streams, in: Proceedings of the International Conference on Process Mining (ICPM 2025) Workshops, 2025.

[13] H. Y. Teh, A. W. Kempa-Liehr, K. I.-K. Wang, Sensor data quality: A systematic review, Journal of Big Data 7 (2020) 11.

[14] A. Tawakuli, B. Havers, V. Gulisano, D. Kaiser, T. Engel, Survey: Time-series data preprocessing: A survey and an empirical analysis, Journal of Engineering Research 13 (2025) 674–711.

[15] M. Shirali, Z. Ahmadi, C. Fernández-Llatas, J. L. Bayo Montón, G. Di Federico, An interactive error-correcting approach for iot-sourced event logs, ACM Transactions on Internet of Things 5 (2024) 1–30.

[16] F. Stertz, S. Rinderle-Ma, J. Mangler, Analyzing Process Concept Drifts Based on Sensor Event Streams During Runtime, in: D. Fahland, C. Ghidini, J. Becker, M. Dumas (Eds.), Business Process Management, volume 12168 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 202–219. doi:10.1007/978-3-030-58666-9_12.

[17] B. Steenwinckel, D. de Paepe, S. Vanden Hautte, P. Heyvaert, M. Bentefrit, P. Moens, A. Dimou, B. van den Bossche, F. de Turck, S. van Hoecke, F. Ongenae, Flags: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning, Future Generation Computer Systems 116 (2021) 30–48. URL: https://www.sciencedirect.com/science/article/pii/S0167739X20329927. doi:10.1016/j.future.2020.

10.015.

[18] S. Ghalibafan, S. Sadeghianasl, M. T. Wynn, eRooMiner: A Data-Driven Approach for Root Cause Detection of Process Data Quality Issues, in: C. Cappiello, O. Hartig, M. Sellami, A. Ouni (Eds.), Cooperative information systems, volume 15535 of *Lecture Notes in Computer Science*, Springer, Cham, 2026, pp. 33–51. doi:10.1007/978-3-032-15538-2_3.

[19] R. Andrews, F. Emamjome, A. H. ter Hofstede, H. A. Reijers, Root-cause analysis of process-data quality problems, Journal of Business Analytics 5 (2022) 51–75. doi:10.1080/2573234X.2021.1947751.

[20] R. Seiger, S. Huber, P. Heisig, U. Aßmann, Enabling Self-adaptive Workflows for Cyber-physical Systems, in: R. Schmidt, W. Guédria, I. Bider, S. Guerreiro (Eds.), Enterprise, Business-Process and Information Systems Modeling - 17th International Conference, BPMDS 2016, 21st International Conference, EMMSAD 2016, Held at CAiSE 2016, Ljubljana, Slovenia, June 13-14, 2016, Proceedings, Lecture Notes in Business Information Processing, Springer, 2016, pp. 3–17. doi:10.1007/978-3-319-39429-9_1.

[21] L. Malburg, M. Hoffmann, R. Bergmann, Applying MAPE-K control loops for adaptive workflow management in smart factories, Journal of Intelligent Information Systems 61 (2023) 83–111. doi:10.1007/s10844-022-00766-w.

[22] A. Bazan-Muñoz, C. Pautasso, G. Ortiz, A. G. de Prado, Bpsmart-care: a framework for managing contextualized actions in iot systems through the integration of business process modelling and complex event processing, Internet of Things 36 (2026) 101887. URL: https://www.sciencedirect.com/science/article/pii/S254266052600017X. doi:https://doi.org/10.1016/j.iot.2026.101887.

[23] A. Schultheis, Y. Bertrand, J. Grüger, L. Malburg, R. Bergmann, E. Serral Asensio, Case-based data quality management for iot logs: A case study focusing on detection of data quality issues, IoT 6 (2025). URL: https://www.mdpi.com/2624-831X/6/4/63. doi:10.3390/iot6040063.

[24] A. Schultheis, B. Alt, S. Bast, A. Guldner, D. Jilg, D. Katic, J. Mundorf, T. Schlagenhauf, S. Weber, R. Bergmann, S. Bergweiler, L. Creutz, G. Dartmann, L. Malburg, S. Naumann, M. Rezapour, M. Ruskowski, EASY: Energy-Efficient Analysis and Control Processes in the Dynamic Edge-Cloud Continuum for Industrial Manufacturing, KI - Künstliche Intelligenz (2024). doi:10.1007/s13218-024-00868-3.

[25] B. Fekade, T. Maksymyuk, M. Kyryk, M. Jo, Probabilistic recovery of incomplete sensed data in iot, IEEE Internet of Things Journal 5 (2018) 2282–2292. doi:10.1109/JIOT.2017.2730360.

[26] R. Seiger, F. Zerbato, A. Burattin, L. García-Bañuelos, B. Weber, Towards iot-driven process event log generation for conformance checking in smart factories, in: 2020 IEEE 24th International Enterprise Distributed Object Computing Workshop (EDOCW), 2020, pp. 20–26. doi:10.1109/EDOCW49879.2020.00016.

[27] E. Brzychczy, U. Jessen, K. Kluza, S. Sriram, M. V. Nettelnstroth, Iot miner: Intelligent extraction of event logs from sensor data for process mining, ArXiv abs/2509.05769 (2025). URL: https://api.semanticscholar.org/CorpusID:281202837.

[28] A. Pery, M. Rafiei, M. Simon, W. M. van der Aalst, Trustworthy artificial intelligence and process mining: challenges and opportunities, in: International Conference on Process Mining, Springer, 2021, pp. 395–407.

[29] K. Moutselos, D. Kyriazis, V. Diamantopoulou, I. Maglogiannis, Trustworthy data processing for health analytics tasks, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 3774–3779.

[30] J. Friederich, S. Lazarova-Molnar, Process mining for reliability modeling of manufacturing systems with limited data availability, in: 2021 8th International Conference on Internet of Things: Systems, Management and Security (IOTSMS), IEEE, 2021, pp. 1–7.

[31] A. Stevens, J. De Smedt, J. Peeperkorn, J. De Weerdt, Assessing the robustness in predictive process monitoring through adversarial attacks, in: 2022 4th international conference on process mining (ICPM), IEEE, 2022, pp. 56–63.

[32] A. Stevens, J. Peeperkorn, J. De Smedt, J. De Weerdt, Manifold learning for adversarial robustness

in predictive process monitoring, in: 2023 5th International Conference on Process Mining (ICPM), IEEE, 2023, pp. 17–24.
[33] A. Maldonado, C. M. M. Frey, G. M. Tavares, N. Rehwald, T. Seidl, Gedi: Generating event data with intentional features for benchmarking process mining, in: Proceedings of Business Process Management (BPM 2024), 2024, pp. 221–237. URL: https://dblp.org/rec/conf/bpm/MaldonadoFTRS24. doi:10.1007/978-3-031-70396-6_13.
[34] A. Maldonado, C. M. M. Frey, S. A. Aryasomayajula, S. A. Fahrenkrog-Petersen, L. Zellner, T. Seidl, Shaining on process mining: Explaining event log characteristics impact on algorithms, in: Proceedings of International Conference on Process Mining (ICPM'25), Montevideo, Uruguay, 2025.
[35] C. Imenkamp, A. Maldonado, H. Reiter, M. Werner, W. Hasselbring, A. Koschmider, A. Burattin, Avocado: The streaming process mining challenge, arXiv preprint arXiv:2510.17089 (2025).
[36] A. Macías, D. Muñoz, E. Navarro, P. González, Data fabric and digital twins: An integrated approach for data fusion design and evaluation of pervasive systems, Information Fusion 103 (2024) 102139. URL: https://www.sciencedirect.com/science/article/pii/S1566253523004554. doi:https://doi.org/10.1016/j.inffus.2023.102139.
[37] M. T. Wynn, R. Andrews, S. Sadeghianasl, Process-Data Quality Management Lifecycle: Building Strong Data Foundations for Process Mining, Springer Nature Switzerland, Cham, 2026, pp. 384–397. URL: https://doi.org/10.1007/978-3-032-17618-9_27. doi:10.1007/978-3-032-17618-9_27.