







# Challenges in Data Quality Management for IoT-Enhanced Event Logs<sup>\*</sup>

Yannis Bertrand<sup>1</sup>, Alexander Schultheis<sup>2,3</sup>, Lukas Malburg<sup>2,3</sup>, Joscha Grüger<sup>2,3</sup>, Estefanía Serral Asensio<sup>4</sup>, and Ralph Bergmann<sup>2,3</sup>

<sup>1</sup> Department of Business Informatics and Operations Management, Faculty of Economics and Business Administration, Ghent University, Ghent, Belgium

<sup>2</sup> German Research Center for Artificial Intelligence (DFKI),  
Branch Trier University, Behringstraße 21, 54296 Trier, Germany

<sup>3</sup> Artificial Intelligence and Intelligent Information Systems, Trier University,  
54296 Trier, Germany, <https://www.wi2.uni-trier.de>

<sup>4</sup> Research Centre for Information Systems (LIRIS), KU Leuven, Brussels, Belgium  
[yannis.bertrand@ugent.be](mailto:yannis.bertrand@ugent.be)  
{alexander.schultheis, lukas.malburg, joscha.grueger, ralph.bergmann}@dfki.de  
{schultheis, malburg1, grueger, bergmann}@uni-trier.de  
[estefania.serralasensio@kuleuven.be](mailto:estefania.serralasensio@kuleuven.be)

**Abstract.** Modern organizations make frequent use of *Internet of Things* (IoT) devices, such as sensors and actuators, to monitor and support their so-called IoT-enhanced *Business Processes* (BPs). These IoT devices collect vast amounts of data which, when processed appropriately, can yield crucial insights into the working of the BPs. However, IoT data, such as sensor data, is notoriously of poor quality, e.g., suffering from noise or having some missing data points. These problems are referred to as *Data Quality Issues* (DQIs), which often interfere with the analysis of IoT data in an industrial context. In this paper, we present a list of challenges that have to be tackled to achieve *Data Quality* (DQ) management in IoT-enhanced event logs. These challenges are derived and refined by leveraging expert knowledge and experience in DQIs within a focus group interview. In addition, we provide directions for solutions to these challenges based on input from the focus group interview and the literature. Finally, we discuss the challenges and their impact on typical DQ management tasks. The insights provided can help guide future research to achieve better DQ in event logs of IoT-enhanced BPs.

**Keywords:** Business Process Management · Internet of Things · Data Quality Management · Data Quality Issues

## 1 Introduction

Recently, the field of *Business Process Management* (BPM) is getting more and more attention in the context of *Internet of Things* (IoT) environments

---

<sup>\*</sup> The final authenticated publication is available online at [https://doi.org/10.1007/978-3-031-92474-3\\_2](https://doi.org/10.1007/978-3-031-92474-3_2)

[15, 19, 33] such as smart homes or smart factories. In this context, BPM can be applied to execute modeled processes with support from IoT devices in smart environments, realizing IoT-enhanced *Business Processes* (BPs). In these processes, IoT sensor data can be used to detect certain emerging situations and react to them [26]. The integration of IoT data and event data generated by BPM systems is not straightforward, and previous research has presented multiple challenges [7, 15]. However, the literature has so far not dealt with an essential hurdle to IoT-enhanced BPM, namely *Data Quality* (DQ) [17]. IoT sensor data are prone to *Data Quality Issues* (DQIs), which can have various root causes, such as sensor failures, network problems, or environmental influences [14, 37].

DQIs can significantly affect both the quality of the data and the results of subsequent analyzes [8]. Previous research has identified patterns that can characterize recurring DQIs in the data. However, there remain crucial challenges to overcome in DQ management in IoT-enhanced BPM. For example, it remains challenging to address DQIs that only rarely occur, but have a significant impact on the DQ. These infrequent issues often require domain specialists for the detection and handling of these DQIs. Moreover, methods and approaches supporting domain experts in this process are currently still in their infancy and only support certain types of DQIs or have a very high computational complexity (see, e.g., Schultheis et al. [28]), leading to low user satisfaction. In total, this gap limits the ability to handle DQIs, leading to poor analysis and decisions based on the event log.

In general, the goal of this paper is to highlight critical challenges in managing DQIs within IoT-enhanced event logs. This paves the way for the development of intelligent techniques to resolve DQIs and ensure the acquisition of high-quality sensor data, which can be analyzed alongside process data provided by a process-aware information system, or abstracted to detect process events in the absence of such a system. Specifically, this paper’s contribution is twofold:

1. It identifies challenges that hamper DQ management in IoT-enhanced event logs, following a mixed-method approach that combines a literature review with a focus group interview; and
2. it examines the relationship between these challenges and the typical DQ management tasks discussed in the literature.

Therefore, this paper is structured as follows: Section 2 provides the background necessary to understand the topic of IoT-enhanced BPM and the challenges in managing DQIs, particularly within IoT event logs, and the main phases of DQ management addressed in the literature. Then, Section 3 introduces the methodology followed to derive the challenges, involving a focus group interview with IoT BP experts. In Section 4, the key challenges associated with addressing DQIs are identified and elaborated. These challenges are positioned within the broader context of DQ management in Section 5, illustrating their relevance to established DQ management tasks. Subsequently, in Section 6 possibilities based on which the various groups of challenges can be addressed are outlined. Section 7 offers a discussion of the implications of these challenges and how they relate

to existing approaches as well as threads to the validity of the results. Finally, Section 8 concludes the paper, summarizing the main findings and suggesting directions for future research.

## 2 Background

This section presents the research background of this work. Section 2.1 presents the state of research in the field of DQ, while Section 2.2 shows the IoT-enhanced research in the BPM area. Section 2.3 covers preliminary work that combines both research areas. The management of DQ, which arises in this context, is introduced in Section 2.4.

### 2.1 Data Quality

DQ is a research area that mainly deals with the detection and handling of DQIs in databases and datasets. In general, DQ is considered to determine the extent to which the data meet the requirements of their users [4, 36]. Various dimensions are defined to describe and quantify DQ, among them: accuracy, timeliness, precision, completeness, reliability, and error recovery [17]. The importance of each of these dimensions depends on the application scenario and the type of data.

### 2.2 IoT-Enhanced Business Process Management

IoT-enhanced BPs are characterized by the multitude of IoT devices (i.e., sensors and actuators) that support their execution, for example, by automating tasks and tracking physical process parameters [15]. Recently, there has been increasing awareness of the potential integration of IoT data with BPM techniques in IoT-enhanced BPs in various sectors, including manufacturing, healthcare, logistics, and smart spaces [3, 10, 19, 29].

There, IoT data can be used to contextualize BPs with data describing their physical environment, e.g., sensor data in modern production processes or patient vital signs in healthcare. For example, IoT-enhanced process mining techniques can provide a more profound understanding of the BPs by, e.g., deriving IoT-based decision rules, process variants, or anomalous sensor data patterns. There, typical steps include preprocessing the raw data (i.e., cleaning, formatting), event correlation to retrieve the cases each event belongs to, and event abstraction to derive meaningful process events from sensor data [10, 18].

### 2.3 Data Quality in IoT Business Process Management

IoT data quality is a broad topic, ranging from detecting DQIs to improving DQ through cleaning methods [17, 34]. IoT applications often rely on low-cost sensors with limited battery and processing power, often deployed in hostile

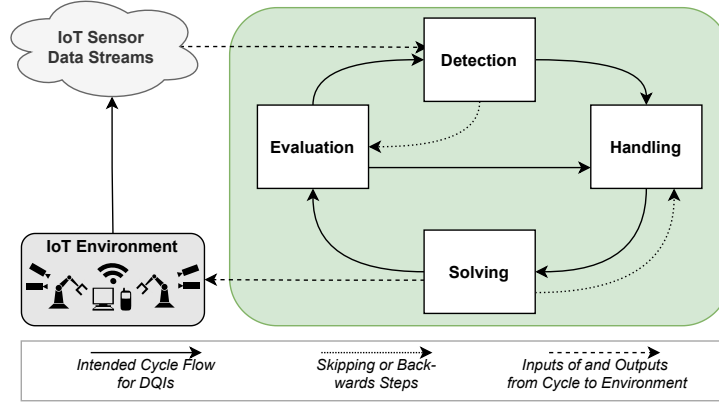
environments [34]. This leads to sensor issues such as low sensing accuracy, calibration loss, sensor failures, improper device placement, range limit, and loss of data packages. Such sensor faults, in turn, cause various types of errors in the generated data, complicating further analysis. Teh et al. [34] review the sensor DQ literature and list the following error types (in decreasing order of frequency): outliers; missing data; bias; drift; noise; constant value; uncertainty; stuck-at-zero. When left untreated, these errors lead to incorrect data and unreliable analysis, leading to poor decision-making.

Although basically all previous research applying BPM with IoT data had to tackle DQIs, the existing body of literature specifically addressing DQIs in IoT-enhanced BPM is still scarce. Bertrand et al. [8] review the IoT BPM literature to describe the IoT DQIs and the event log DQIs encountered by previous research. Based on this state of affairs, patterns that relate IoT DQIs with resulting event log DQIs are derived. More specifically, these patterns are of the following shape: Sensor fault  $\implies$  sensor DQI  $\implies$  event log DQI(s). For example, one of the patterns describes the following: unstable environment  $\implies$  noisy sensor data  $\implies$  incorrect case ID. A concrete example of the occurrence of this pattern in the literature is provided by Brzychczy and Trzcionkowska [10], where data from sensors placed on a drilling machine in a mine were used to derive an event log. Unfortunately, the sensors produced noisy data, making it difficult to recognize the start and end activities of the mining process and resulting in some incorrect case IDs in the log derived from the sensor data.

## 2.4 Data Quality Management

Various steps for addressing DQIs are discussed in the literature (e.g., in [5, 12, 14, 34]). In this section, we introduce some of the most commonly discussed tasks in DQ management. Although so far a widely accepted standardized model has not yet been proposed that describes the process of DQ management from detection to handling of DQIs, we outline the main tasks in Figure 1.

- 1. Detection:** In this task, incoming IoT sensor data are monitored to identify possible DQIs [14, 34]. The result of this is the classification of whether DQIs are present and, if so, which ones. If DQIs are detected, the data are passed on to the handling phase.
- 2. Handling:** This task typically builds on the classification of DQIs output by the detection [14]. The aim of handling is to provide recommendations for action to rectify each individual DQI. These may include both the repair of the components that caused the DQI and the cleaning of the event log. As a result, recommendations for action are given for the solving task.
- 3. Solving:** In this task, scheduling is carried out to prioritize the recommended actions from handling [12]. For example, duplicate steps are combined, and actions are checked for interdependencies. The actions are then implemented automatically, semi-automatically or manually [5, 34]. If it turns out that the proposed action cannot be executed, the problem is returned to the handling task.



**Fig. 1.** Visualization of the Relationships Between Common DQ Management Steps (In Accordance With the Plan-Do-Check-Act Cycle [21])

**4. Evaluation:** The evaluation verifies the correctness of the result, i.e., whether the log is cleaned and/or the cause of the DQIs is solved [12, 14]. This task should also check whether new DQIs have been generated as a result of the application of a data cleaning technique. If insufficient quality is detected in the evaluation, it is checked whether the identification and classification of the DQI in the detection phase is correct and whether the recommended and executed actions have really solved this problem. If this is not the case, the detection and handling tasks can be triggered again.

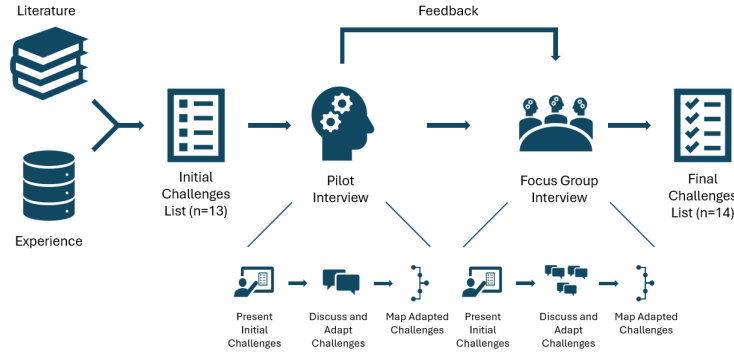
### 3 Focus Group Methodology

In this section, we introduce the mixed method approach we used to derive the challenges (depicted in Figure 2).

First, we created an initial list of challenges, derived from two main sources: 1) the challenges identified in the literature [8] (C1-C5 as described in Section 4); and 2) challenges the authors faced in their work with IoT-enhanced business processes [7, 29, 28] (C6-C7, and C9-C14 as described in Section 4).

Afterward, we carried out a pilot interview with an expert in IoT-enhanced processes to prepare and test the focus group interview. The feedback obtained from the pilot interview was used to refine the interview process. Notable improvements that have been made following the pilot interview include a greater focus on qualitative feedback about the challenges and the grouping of the challenges.

Then, we conducted a focus group interview [24]. The focus group interview method involves selected participants discussing a specific topic, using group dynamics to produce more in-depth insights than individual interviews [35]. Participants are chosen for their relevance and expertise to the topic and their



**Fig. 2.** Overview of the Methodology Followed to Derive the Challenges Presented in the Paper.

previous experience in group discussions [25]. We selected four researchers who are specialized in IoT-enhanced BPM, all members of the *Internet of Processes and Things* (IoPT) initiative<sup>5</sup>, a group of researchers pioneering industrial IoT applications. As we were able to consult these members of the IoPT group, we gained detailed insights and challenges in current research works on the topic of IoT-enhanced BPM through the interview. In addition to their research work, this also includes the practice-oriented research projects the members are conducting with industry partners, which go beyond laboratory experiments and, thus, provide more realistic and industrial challenges. The focus group has been designed to facilitate discussion and stimulate interaction between participants. Specifically, the participants were first presented with the initial list of challenges. The participants have then been asked to discuss the challenges, with the discussion revolving around three main questions:

1. Which challenges have they already encountered?
2. Which challenges do they consider irrelevant?
3. Which challenges do they believe are missing?

During the interview, the group evaluated, adapted, and grouped the proposed challenges, and multiple challenges were proposed and debated. Challenges have only been included by the focus group if a consensus has been reached, and the focus group ultimately resulted in the inclusion of one more challenge in the final list (C8). Meanwhile, the authors moderated the exchanges, pointing out some similarities and differences between the proposals made by the participants. The discussion has been recorded and transcribed for analysis<sup>6</sup>. This group interview resulted in a general consensus on the final list of challenges presented in Section 4.

Throughout the discussion, suggestions have been made to solve the challenges. These suggestions are summarized and completed in Section 6, where,

<sup>5</sup> <https://zenodo.org/communities/iopt/about>

<sup>6</sup> Transcript available at: <https://zenodo.org/records/15058151>

for each group of challenges, we present potential solutions and approaches to mediate the impact of the challenges. Finally, the participants were asked to reflect upon which task(s) of DQ management presented in Section 2.4 each group of challenges impacted most, resulting in the mapping presented in Section 5.

## 4 Challenges for Addressing Data Quality Issues

In this section, we present the list of challenges derived from the focus group interview. In total, 14 challenges were identified, which are clustered into four groups: Data Complexity (see Section 4.1), Knowledge Acquisition (see Section 4.2), Knowledge Management (see Section 4.3), and Conflict of Interests (see Section 4.4).

### 4.1 Group 1: Data Complexity

This group addresses the challenges related to the fundamental characteristics of big data in the context of IoT environments. Specifically, it focuses on issues that arise from the immense volumes of data generated by sensors, the variety of data formats, the velocity at which data must be processed, and the granularity required for effective DQI identification and resolution. The following four challenges are identified for this group:

**C1 – Volume:** This challenge pertains to the big data issue of volume [16], which involves the continuous generation and collection of vast amounts of data. In particular, IoT BPs equipped with numerous sensors, especially in long-running operations, produce an overwhelming volume of data. Analyzing these large data sets is very computationally intensive [28]. Although approaches that simplify or abstract data representation (e.g., [2, 20, 31]) can reduce computational load, they often result in a trade-off between data compression and the loss of critical information, potentially causing some DQIs to go undetected.

**C2 – Variety:** Variety refers to the big data challenge of managing data produced in multiple formats – structured, unstructured, or semi-structured – from diverse sources [16]. In the context of DQIs, this variety emerges from different types of sensors that generate time series data, which represent various values over time, such as Boolean states or spatial coordinates [2]. Additionally, the complexity increases with dependencies between data sources, requiring careful consideration in data storage and processing.

**C3 – Velocity:** Velocity refers to the challenge of processing data at high speed, as described in the context of big data [16]. In IoT BPs, sensors perform frequent measurements, typically every second. To effectively address DQIs, it is crucial to identify and correct failures in near real time. Delaying this process can result in a trade-off where failures are only analyzed after they have impacted the data, compromising the timeliness of DQI management.

**C4 – Granularity:** Granularity addresses the level at which DQIs are identified and solved. The key is to choose the appropriate method or algorithm for taking the data to the desired level of detail. For instance, it can be advantageous to pre-process or abstract data (see the trade-off in C1) instead of relying on raw sensor data. The granularity level enables different processing places, such as distributed pre-processing at the edge level [27, 32]. However, data abstraction is performed at the risk of overlooking dependencies, such as those between time series, which may only be fully captured at lower levels of granularity.

## 4.2 Group 2: DQI Knowledge Acquisition

To effectively identify and handle DQIs, it is crucial to acquire knowledge about them. The challenges in this category are associated with the collection and processing of knowledge regarding the DQIs present in the data, which is closely aligned with the research area of knowledge acquisition [23]. This process pertains to the specific methods of extracting knowledge from the data. Therefore, four key challenges are identified.

**C5 – New Data Quality Issues:** This challenge deals with DQIs that have never been observed before. Hence, there cannot be knowledge about this type of failure, making automatic detection and classification difficult. This raises the problem of how knowledge about such new DQIs can be suitably acquired and generated, e.g., for automated processing in an Artificial Intelligence model. Although there are methods to identify the existence of an unknown issue, such as anomaly detection [11], concrete classification requires further information.

**C6 – Rare Data Quality Issues:** Rare DQIs present significant challenges due to the scarcity of data on these infrequent occurrences. This scarcity complicates pattern recognition, algorithm training, and strategy development to manage them. Classifying such DQIs and obtaining relevant information through appropriate solutions is particularly difficult. Moreover, there is a heightened risk of, e.g., overfitting when training Machine Learning models to detect these rare DQIs. The core challenge, therefore, lies in effectively addressing rare DQIs throughout the various steps of DQ management described in Section 2.4, despite the limited and imbalanced available data.

**C7 – Causality Ambiguity:** In the context of DQIs, ambiguity means that known patterns cannot be clearly traced back to an issue. For example, noisy data can result from a sensor being incorrectly calibrated or from an issue in the environment, such as someone walking near the sensor. In this case, the DQI patterns left in the data would be very similar, but the cause would be entirely different, creating difficulties to correctly classify and handle the DQI.

**C8 – Solution Ambiguity:** This challenge involves the difficulty of selecting the appropriate solution for a DQI when ambiguity exists. Even if the cause of a DQI is known, there may be multiple solutions with no clear best choice. This uncertainty creates challenges in evaluating and applying different possible remedies. Implementing a solution might trigger a conflict of business interests



(see C13), such as requiring significant financial investment or resources. Furthermore, there might be dependencies between solutions that could introduce new DQIs (see C14). The key issue is effectively balancing and choosing among various solutions, each with its own trade-offs and potential implications.

### 4.3 Group 3: DQI Knowledge Management

In this section, we explore the challenges related to the management of existing knowledge about DQIs. Effective knowledge management [6, pp. 9-11] is essential for the identification, handling, and resolution of DQIs, especially when dealing with complex and evolving data environments. The challenges discussed here focus on the representation, reuse, and refinement of knowledge. These aspects are crucial to ensure that knowledge about DQIs is not only well-documented and structured, but also effectively leveraged and continuously improved to address both existing and emerging issues in DQ management.

**C9 – Knowledge Representation:** This challenge addresses the problem of finding a suitable representation of knowledge. It considers the preprocessing of existing knowledge about DQIs (e.g., manuals, workers’ experience, patterns, causes, and solutions) into a format interpretable for machines so that it can be processed automatically (so-called externalization [22]). The appropriate knowledge representation is the basic requirement for applying algorithms or more advanced methods to address DQIs.

**C10 – Knowledge Reuse and Transfer:** Knowledge Reuse and Transfer involves leveraging existing, well-structured knowledge stored in a knowledge base (e.g., a database) to detect, manage, and solve new DQIs. The key challenge is to automatically identify the most relevant knowledge within the database and adapt it to address new DQIs. This process can occur within the current environment (e.g., factory or machine) or across different environments, which may require further generalization to ensure effective application.

**C11 – Knowledge Refinement:** This challenge refers to the constant evaluation and, if necessary, updating of knowledge. For example, solutions that are proposed for a new DQI as part of knowledge management and applied where necessary must be checked to see whether they actually increase the DQ for this use case. Furthermore, it is possible that the application of solutions may cause new problems (see C14). To ensure that a DQ management system, that is as automated as possible, eliminates causes of DQIs and increases DQ, constant refinement of the knowledge must therefore be carried out. In this context, knowledge might also be acquired again, referring to the challenges of Group 2 (see Section 4.4).

### 4.4 Group 4: Conflict of Interests

This section addresses the challenges that arise when DQ management concerns clash with broader organizational matters. These include issues such as privacy, where data protection requirements can impede the identification and resolution

of DQIs; business interest conflicts, where financial considerations may restrict the resources dedicated to DQ management; and cascading effects, where the resolution of DQIs might unintentionally cause adverse consequences in the data processing workflow. Addressing these challenges is essential for implementing effective and sustainable DQ management practices.

**C12 – Privacy:** This challenge highlights potential conflicts of interest related to data privacy, which can occur in two main situations. First, privacy may be required to protect personal interests or to comply with legal regulations. For instance, these privacy concerns may conflict with the analysis of certain data, such as video recordings, or may entirely prohibit the use of whole data sets and data sources. This means that data that could be used to identify and rectify DQIs may be missing, or their use could be restricted. On the other hand, privacy can also be viewed in terms of trade secrets and confidentiality, where sensitive information is protected from unauthorized access or disclosure. This means that a cross-enterprise exchange of possible DQI knowledge may be restricted to prevent competitors from obtaining information about production (which exacerbates the problem of few failure data, see C6).

**C13 – Business Interest Conflicts:** This challenge describes financial interests within a company that may stand in the way of DQ management. For example, this DQ management may play a subordinate role within the company and there may be no willingness to spend money on addressing it at all, e.g., by deciding not to replace a faulty sensor. Certain DQIs may also be considered unimportant and contrary to other internal company guidelines. For example, Predictive Maintenance [38] is often carried out within a factory. If it is recognized that a sensor is producing faulty or missing data, it may still be economical from this perspective to keep the sensor running shortly before or even until total failure. On the other hand, DQ management would have an interest in repairing or even replacing the sensor as fast as possible.

**C14 – Cascading Effects:** This challenge can occur if the semantics of a process can get lost due to the elimination of DQIs. In general, there is an interest in rectifying DQIs if possible. However, this can lead to undesirable side effects, as a result of which the data are still not suitable for analysis, its quality may even deteriorate or non-existent other failures may be searched for. For example, a faulty sensor might disturb the production in a smart industrial process. By repairing the sensor causing the failure and the corresponding DQI, the process can be continued. However, repairing the event log would mean that the semantics that caused production to stop would no longer be retained. As a result, the event log would be searched for the failure that caused the stop and none would be found, leading to an unnecessary failure investigation. In this case, the interest of a clean event log takes precedence over the semantics of the process. Such cases must therefore be identified and handled accordingly, possibly by discarding a part of the log.

**Table 1.** Mapping of Challenges to the Four Main DQ Management Tasks Presented in Section 2.4.

Group	Detection	Handling	Solving	Evaluation
<b>Group 1</b> (Section 4.1)	✗	✗		
<b>Group 2</b> (Section 4.2)	✗	✗	✗	✗
<b>Group 3</b> (Section 4.3)	✗	✗	✗	✗
<b>Group 4</b> (Section 4.4)	✗	✗	✗	

## 5 Mapping of Challenges to Data Quality Management Tasks

The final part of the focus group interview aimed at mapping the challenges with the DQ management tasks discussed in Section 2.4. This positions the challenges in the broader context of DQ management. Depending on the addressed step, it is possible to determine which challenges relevant to this step are likely to occur and based on this, define a prioritization for dealing with them. Depending on the step, various solution approaches can be developed, each addressing the most relevant challenges. This prioritization was derived and evaluated through discussions with the experts in the focus group interview. As the participants saw little difference between the challenges within each group, the mapping is given for each group of challenges. The mapping is shown in Table 1.

First, for Group 1, it was recognized that challenges C1-C4 mostly impact the detection and the handling of the DQIs, as these tasks usually assume a simplification of the data which negates the effect of the challenges on subsequent tasks. For example, with large volumes of data (C1), the share of data where DQIs are detected is likely to be minimal, which means that the volume of data that will go through the next tasks will typically be more manageable. Then, for Group 2, all tasks are impacted: in C5 and C6, the rarity of the DQI implies difficulties in detection as well as solving and the evaluation of cleaned data, while ambiguity can affect both the detection and evaluation tasks (C7) as well as the handling and solving phases (C8). Next to this, the Group 3 challenges also impact all tasks, as the knowledge about the DQIs covers the techniques to detect them as well as the approaches to solve them and their effects. Finally, Group 4 affects all tasks but the evaluation, as these challenges relate to the availability of data (especially C12) and the prioritization of the solutions that can be applied (C13 and C14).

## 6 Towards Addressing the Challenges

In this section, we discuss possible solutions to the challenges based on the results of the focus group interview, completed with the literature. We present the solutions along the four groups of DQIs identified in Section 4.

### 6.1 Group 1: Data Complexity

The challenges in this group are deeply interrelated and are best addressed all together. For instance, more fine-granular data (C4) and different types of data (C2) often lead to a higher volume of data (C1), which in turn exacerbates the challenge of velocity (C3). Overall, a trade-off has to be made between them: To tackle the challenge of velocity and solve DQIs faster, one can, e.g., abstract the data to a higher granularity level, thereby also reducing the volume of data, making it easier to process the data in a timely way (albeit at the cost of potentially missing some DQIs).

Two primary approaches were proposed to tackle these challenges holistically:

- a) *Expert-Guided Data Prioritization & Filtering*: By focusing on the most relevant bits of data and aggregating data to a semantically meaningful level, the volume of data to process can be reduced, making it easier to analyze various types of data (C2), timely, and at the right granularity level. This technique can be supported by a divide-and-conquer approach to data preprocessing, where an initial layer of data filtering occurs at the edge, streamlining subsequent processing.
- b) *Increasing Computational Power*: Expanding processing capacity enables the handling of larger data volumes (C1), faster processing of the same amount of data (C3), analysis at a finer level of granularity (C4), or a combination of these improvements simultaneously.

### 6.2 Group 2: DQI Knowledge Acquisition

For the challenges in group 2, potential solutions lie in specific data analysis techniques and expert knowledge. To address new DQIs (C5), anomaly detection techniques can be employed to identify new types of DQIs (for an in-depth treatment of anomaly detection in IoT time series, see [11, 30]). Next to this, expert knowledge can guide the discovery of new DQI types, in particular when the process is being modified, i.e., there is concept drift. Then, for rare DQIs (C6), data augmentation techniques can be used to generate more cases of emerging DQI types to refine the DQI detection and handling techniques. Next, to address causality ambiguity (C7), root cause analysis based on expert knowledge is the most straightforward option. This way, occurrences of similar DQIs due to different causes could be more easily disentangled. Solution ambiguity (C8), in turn, can be addressed by ranking solutions and trying to apply the best ranked solutions until the DQI is solved. This ranking could be performed based on the likelihood that each solution will solve the issue and its cost.

### 6.3 Group 3: DQI Knowledge Management

This group of challenges revolves around knowledge representation, and the choice of a suitable knowledge representation, addressing knowledge representation (C9), is a prerequisite to tackling the other challenges. To this end, we expect

a suitable knowledge representation to store, for each DQI detected, the data, the cause of the DQI and the solution applied. Knowledge-based approaches (like case-based reasoning [1, 6]), ontologies [9] or knowledge graphs [13] could be used to store this type of information. Moreover, they are supported by data retrieval techniques that can be applied to address knowledge reuse and transfer (C10). Knowledge refinement (C11), finally, can also be addressed by the techniques used for new DQIs (C5) and rare DQIs (C6), as the detection of new DQI types is necessary to the update of the knowledge on these DQIs.

#### 6.4 Group 4: Conflict of Interests

The challenges in this group are not as interconnected as the challenges in other groups and require different solutions. Depending on the type of data and the level of confidentiality required, data privacy (C12) can be addressed by anonymizing the data, using federated learning to learn from multiple data sources without directly sharing the data or generating synthetic data. For business interest conflicts (C13), the issue is mainly a management one, which requires prioritizing DQIs based on a cost-benefit analysis, and only solving DQIs which cause more financial damage than the cost of handling them. Regarding cascading effects (C14), one possibility is to keep both the original log (containing DQIs) and the cleaned one, so that root cause analysis is always possible on the original data which caused the issues in the process execution.

## 7 Discussion

One of the main difficulties encountered in the collection and description of the challenges presented in Section 4 is that such challenges are very interdependent. This interdependence is often observed between challenges of the same group. For example, granularity (C4) and volume (C1) impact each other, as being able to abstract the raw sensor data without losing important information can greatly reduce the burden caused by large volumes of data. Moreover, solution ambiguity (C8) can be mildewed by accurate detection of DQIs, i.e., resolving causality ambiguity (C7).

Some challenges can also have an effect on other challenges in other groups, e.g., the granularity level at which data are analyzed (see C4) can be impacted by the knowledge representation (C9) chosen. As another example, a suitable procedure for business interest conflicts (C13) can mitigate rare DQI challenges (see C6).

Next to this, some overarching factors that can affect multiple challenges are identified. First, in the handling and solving phases, two types of techniques can be proposed and applied:

1. Approaches to clean the data, e.g., scripts to remove outliers or impute missing data,
2. Actions to repair the data collection process, e.g., recalibrating a sensor or replacing a faulty network component.

These have very different effects and can solve different types of issues, as repairing the data collection process is typically more costly (e.g., regarding hardware costs or downtimes) and takes longer. But some DQIs can only be solved by such interventions. There, business interest conflicts (C13) play a crucial role in balancing the costs and benefits of both types of solutions.

Second, the timeliness of the DQ management (i.e., how fast DQIs can and need to be solved), while having an obvious link with the velocity challenge (C3), also impacts multiple other aspects of DQ management and determines how serious some challenges are. As such, the volume of the data (C1) can be handled more easily if the data can be stored and processed per batch. The challenges of Group 2 (see Section 4.2) are also easier to address if there is enough time to have a process/data expert analyzing the problem in depth. On the other hand, being able to address DQIs in (near) real-time makes some challenges much less problematic. For example, cascading effects (C14) can be defused before they have a chance to occur if the initial DQI is immediately handled.

Finally, a last point of discussion in the focus group has been the importance or gravity of each challenge. There, no consensus could be achieved, and it has been suggested that the importance of each challenge is highly dependent on the perspective taken. For example, considering the challenges from a Process Mining perspective, the granularity of the data (C4) has been mentioned as a particularly important challenge.

While these insights provide valuable perspectives on the identified challenges, it is also important to consider potential limitations of our findings. Therefore, we acknowledge the following threats to the validity of our results:

- a) *Limited Sample in the Focus Group Interview*: The challenges were validated by a group of four researchers only as well as one additional researcher for the pilot interview. These focus group participants were selected for their expertise in the domain, which allowed the authors to gather high-quality feedback to validate the challenges.
- b) *Purely Academic Setting*: All participants in the study (authors and focus group interviewees) are university or research institute members. However, all participants have experience working with industry partners, which makes them familiar with the challenges faced when handling real-life data.
- c) *Non-formal Validation Approach*: In this study, we chose the focus group for its ability to generate new ideas through the interaction between participants. However, a focus group is not a systematic validation method, and a controlled survey would allow a more robust validation of the challenges.

## 8 Conclusion and Future Work

In this paper, we address the issue of DQ in the research topic of IoT meets BPM. Our main contribution lies in the discussion of the main challenges that have to be overcome to achieve DQ management in IoT BP, which were derived from

a focus group interview with experts in IoT BPM. Next to this, we propose a mapping of the different challenges groups with the main DQ management tasks discussed in the literature. We put forward the summary of the challenges that we present as a foundation for future research in the domain, and also present initial ideas for addressing these.

In future works, we plan to build on this foundation and develop a framework for semi-automated DQ management in IoT BPs, capable of addressing the challenges we presented, based on the main steps described in Section 2.4.

**Acknowledgments.** This work is supported by the Flemish Fund for Scientific Research (FWO) with grant number G0B6922N. This work is also partially funded by the Federal Ministry for Economic Affairs and Climate Action under grant No. 01MD22002C *EASY* [27].

This research was supported by the Internet of Processes and Things (IoPT) community: <https://zenodo.org/communities/iopt/about>.

## References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Commun.* **7**(1), 39–59 (1994)
2. Allen, J.F.: Maintaining Knowledge about Temporal Intervals. *Commun. ACM* **26**(11), 832–843 (1983)
3. Banham, A., Leemans, S.J., Wynn, M.T., Andrews, R., Laupland, K.B., Shinnars, L.: xPM: Enhancing exogenous data visibility. *AI Med.* **133**, 102409 (2022)
4. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. DCSA, Springer (2006)
5. Batini, C., Scannapieco, M.: Data and Information Quality. *Data-Cent. Syst. Appl.* p. 63 (2016)
6. Bergmann, R.: Experience Management: Foundations, Development Methodology, and Internet-Based Applications, LNCS, vol. 2432. Springer (2003)
7. Bertrand, Y., De Weerd, J., Serral, E.: Assessing the suitability of traditional event log standards for iot-enhanced event logs. In: *BPM 2022*. pp. 63–75. Springer (2022)
8. Bertrand, Y., Van Belle, R., De Weerd, J., Serral, E.: Defining Data Quality Issues in Process Mining with IoT Data. In: *ICPM 2022 Workshops. LNBIP*, vol. 468, pp. 422–434. Springer (2022)
9. Breitman, K.K., Casanova, M.A., Truszkowski, W.: Semantic Web: Concepts, Technologies and Applications. NASA MSSE, Springer (2007)
10. Brzychczy, E., Trzcionkowska, A.: Creation of an Event Log From a Low-Level Machinery Monitoring System for Process Mining Purposes. In: *IDEAL 2021*. pp. 54–63. Springer (2018)
11. Cook, A.A., Mısırlı, G., Fan, Z.: Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet Things J.* **7**(7), 6481–6494 (2020)
12. Ehrlinger, L., Wöß, W.: Automated data quality monitoring. In: *22nd ICIQ Proc.* pp. 15.1–15.9. Little Rock, AR (2017)
13. Fensel, D., Simsek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., Wahler, A.: Knowledge Graphs - Methodology, Tools and Selected Use Cases. Springer (2020)

14. Goknil, A., Nguyen, P., Sen, S., Politaki, D., Niavis, H., Pedersen, K.J., Suyuthi, A., Anand, A., Ziegenbein, A.: A Systematic Review of Data Quality in CPS and IoT for Industry 4.0. *ACM Computing Surveys* **55**(14s), 1–38 (2023)
15. Janiesch, C., Koschmider, A., Mecella, M., Weber, B., Burattin, A., Ciccio, C.D., Gal, A., Kannengiesser, U., Mannhardt, F., Mendling, J., Oberweis, A., Reichert, M., Rinderle-Ma, S., Song, W., Su, J., Torres, V., Weidlich, M., Weske, M., Zhang, L.: The Internet-of-Things Meets Business Process Management. A Manifesto. *IEEE Syst. Man Cybern. Mag.* **6**(4), 34–44 (2020)
16. Kaisler, S.H., Armour, F., Espinosa, J.A., Money, W.: Big Data: Issues and Challenges Moving Forward. In: 46th HICSS Proc. pp. 995–1004. IEEE Computer Society (2013)
17. Karkouch, A., Mousannif, H., Al Moatassime, H., Noel, T.: Data Quality in Internet of Things: A state-of-the-art survey. *JNCA* **73**, 57–81 (2016)
18. Koschmider, A., Janssen, D., Mannhardt, F.: Framework for Process Discovery from Sensor Data. In: EMISA. pp. 32–38 (2020)
19. Leotta, F., Mecella, M., Mendling, J.: Applying Process Mining to Smart Spaces: Perspectives and Research Challenges. In: CAiSE 2015 Workshops. pp. 298–304. Springer (2015)
20. Malburg, L., Schultheis, A., Bergmann, R.: Modeling and Using Complex IoT Time Series Data in Case-Based Reasoning: From Application Scenarios to Implementations. In: 31st ICCBR Workshop Proc. vol. 3438, pp. 81–96 (2023)
21. Moen, R., Norman, C.: Evolution of the PDCA cycle (2006)
22. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York (1995)
23. Olson, J.R., Rueter, H.H.: Extracting expertise from experts: Methods for knowledge acquisition. *Expert Syst.* **4**(3), 152–168 (1987)
24. Rabiee, F.: Focus-Group Interview and Data Analysis. *PNS* **63**(4), 655–660 (2004)
25. Richardson, C.A., Rabiee, F.: A Question of Access: An exploration of the factors that influence the health of young males aged 15 to 19 living in Corby and their use of health care services. *Health Educ. J.* **60**(1), 3–16 (2001)
26. Rinderle-Ma, S., Mangler, J.: Process Automation and Process Mining in Manufacturing. In: BPM 2021, LNCS, vol. 12875, pp. 3–14. Springer (2021)
27. Schultheis, A., Alt, B., Bast, S., Guldner, A., Jilg, D., Katic, D., Mundorf, J., Schlagenhauf, T., Weber, S., Bergmann, R., Bergweiler, S., Creutz, L., Dartmann, G., Malburg, L., Naumann, S., Rezapour, M., Ruskowski, M.: EASY: Energy-Efficient Analysis and Control Processes in the Dynamic Edge-Cloud Continuum for Industrial Manufacturing. *KI* (2024)
28. Schultheis, A., Malburg, L., Grüger, J., Weich, J., Bertrand, Y., Bergmann, R., Serral Asensio, E.: Identifying Missing Sensor Values in IoT Time Series Data: A Weight-Based Extension of Similarity Measures for Smart Manufacturing. In: 32nd ICCBR Proc. LNCS, vol. 14775, pp. 240–257. Springer (2024)
29. Seiger, R., Malburg, L., Weber, B., Bergmann, R.: Integrating Process Management and Event Processing in Smart Factories: A Systems Architecture and Use Cases. *J. Manuf. Syst.* **63**, 575–592 (2022)
30. Sgueglia, A., Di Sorbo, A., Visaggio, C.A., Canfora, G.: A systematic literature review of iot time series anomaly detection solutions. *Future Generation Computer Systems* **134**, 170–186 (2022)
31. Shahar, Y.: A Framework for Knowledge-Based Temporal Abstraction. *Artif. Intell.* **90**(1-2), 79–133 (1997)



32. Shi, W., Dustdar, S.: The Promise of Edge Computing. *Computer* **49**(5), 78–81 (2016)
33. Soffer, P., Hinze, A., Koschmider, A., Ziekow, H., Di Ciccio, C., Koldehofe, B., Kopp, O., Jacobsen, A., Sürmeli, J., Song, W.: From Event Streams to Process Models and Back: Challenges and Opportunities. *Inf. Sys.* **81**, 181–200 (2019)
34. Teh, H.Y., Kempa-Liehr, A.W., Wang, K.I.: Sensor data quality: a systematic review. *J. Big Data* **7**(1), 11 (2020)
35. Thomas, L., MacMillan, J., McColl, E., Hale, C., Bond, S.: Comparison of focus group and individual interview methodology in examining patient satisfaction with nursing care. *Soc. Sci. Med.* **1**(4), 206–220 (1995)
36. Wang, R.Y., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. *JMIS* **12**(4), 5–33 (1996)
37. Zhang, L., Jeong, D., Lee, S.: Data Quality Management in the Internet of Things. *Sensors* **21**(17), 5834 (2021)
38. Zonta, T., Da Costa, C.A., da Rosa Righi, R., de Lima, M.J., da Trindade, E.S., Li, G.P.: Predictive maintenance in the Industry 4.0: A systematic literature review. *Comput. Ind. Eng.* **150**, 106889 (2020)