Sensors & Transducers, Vol. 251, Issue 4, April 2021, pp. 37-46

Sensors & Transducers

Published by IFSA Publishing, S. L., 2021
http://www.sensorsportal.com

Data Quality: Enterprise Initiatives' Issues and WSN Challenges

[1, *] Ina Naydenova, [1, 2] Zlatinka Kovacheva and [3] Kalinka Kaloyanova

[1] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Acad. Georgi Bonchev Str., Block 8, 1113 Sofia, Bulgaria
[2] University of Mining and Geology, Sofia, Bulgaria
[3] Sofia University 'St. Kliment Ohridski', Faculty of Mathematics and Informatics
[1] Tel.: (+359 2) 979-38-28, fax: (+359 2) 971-36-49
E-mail: naydenova@gmail.com

Received: 26 March 2021 /Accepted: 28 April 2021 /Published: 30 April 2021

Abstract: This paper summarized the results of our previous work, outlining some important aspects of the data quality provisioning at the organizational level that are usually poorly understood or underestimated and therefore represent a significant setback in compiling a successful data improvement strategy. These aspects do not cover the specifics of the data quality issues in the field of sensor networks. In the current contribution, we examine the major challenges of providing high quality data collected by wireless sensor networks.

Keywords: Data quality, Sensor networks, Challenge, Context, Data quality initiatives, Decision-making.

1. Introduction

Nowadays data supports many enterprise level businesses, forming one of their most valuable assets. Its use can have short-term impact on the businesses but also medium or long-term implications. The way data is being currently used in data analysis to support decision making, including higher-level decisions such as operational (e.g. at a department, or local store level) or tactical (i.e., from the organization point-of-view) is a clear example of such implications. Thus, the need for organizations to possess high quality data is essential, as it can increase the opportunity for the organizations to deliver reliable services [1]. Data quality (DQ) is also a pre-condition for data analytics and Artificial Intelligence (AI) projects. Poor data quality can cause analytics and AI projects to take longer than expected (around 40 % longer), which means they will cost more or even they will eventually fail to achieve the desired results (70 % of AI projects). Therefore, if a company is looking to invest in AI, it must first develop, define, and implement an excellent data quality strategy [2].

The initiatives for provisioning of better data quality have their own features and specifics, in accordance with the different goals and levels of implication they are launched for. In [3], we have addressed several important aspects of enterprise initiatives to provide better data quality for tactical decision-making. In the current contribution, we will address the challenges of providing quality data collected by wireless sensor networks (WSN).

The rest of the paper is organized as follows. In Section 2, we summarize the results of our work given in [3]. We outline some common misunderstandings and underestimated aspects of the data quality provisioning at the organizational level that hinder the establishment of a successful strategy to improve data for the needs of decision making and business intelligence activities. These aspects do not cover the specific characteristics in providing quality data collected by sensor networks. They are focused mainly on issues related to the data quality initiatives' methodology, organization of work, attitude and maturity of the organizations to launch DQ programs, rather than specific technical challenges. Despite that, we believe that issues in discussion make sense for any DQ project, including the initiatives to provide better data for WSN applications. In our experience, a poor understanding of the addressed aspects is one of the main reasons for the failure or unsatisfactory implementation of the DQ initiatives, regardless of the company business activities. However, the WSN applications have their own specific challenges. Therefore, in Section 3, we take a closer look at the WSN-specific data quality issues. Finally, we make some conclusions in Section 4.

2. Misunderstood Aspects of Data Quality Initiatives

We can summarize the following key points that should be taken into account by the organizations looking to improve the quality of their data:

2.1. Contextual Nature of Data Quality

The concept that some data is good ("high quality") for some purposes but is poor ("low quality") for others is not always well assimilated. Incorrect data may have a different impact on the business, depending on the purposes for which the data is used. For example, if a wrong first customer name is registered in an IT system, this indicates the presence of poor data. However, for the task of mail notifications, where only the customer's last name is used, there is no quality issue with the system' data. At the same time, the wrong first name could be a problem if the person's identity needs to be verified. Poor data is often associated with the discrepancies between reality and the data stored in the system, but the quality of data is a much broader concept. Furthermore, data inaccuracies could be an issue for a given use case, but not interfere with others. A similar observation was shared back in 1990 in [4]. The perception that data quality should be defined in accordance with the consumer's needs is further developed in [5]. High quality data can refer to whether data meets the expectations of the users [6], which can be human users or systems, or can be defined as "data that is fit for use by data consumers" [5]. In the ISO/IEC 25012 standard [8] it is defined as "the degree to which a set of characteristics of data fulfills requirements". The authors of [5] also define a "data quality dimension" as a set of data quality attributes that represent a single aspect or construct of data quality. In subsequent studies scholars propose a variety of data quality dimensions. The majority of the authors include some versions of accuracy, completeness, consistency and timeliness among them [6]. It should be noted that quality dimensions give some structure to data quality expectations, but they have nothing to do with the task-dependent nature of the data quality. Let us take the completeness, for example: for the task of mail notification we need a complete and accurate mail address, but for the purposes of geographic segmentation only the city part of the address is required [7].

The contextual nature of the quality is a root of many challenges. The lack of understanding of the duality of data quality (true reflection of reality and dependence on the case of use) leads to a number of problems, such as inappropriate approach in setting data quality goals, difficulties in data quality issues assessment, incorrect expectations regarding the DQ provisioning processes and the business value to be achieved. It also raises the question of the usability of solutions that accumulate large amounts of data from various sources, without clear vision about their purposes of use.

2.2. Task-dependent Objective Measurement

How good is a company's data quality? Answering this question requires usable data quality metrics. Companies must deal with both the subjective perceptions of the individuals involved with the data, and the objective measurements based on the data set in question. Subjective data quality assessments reflect the needs and experiences of stakeholders: the collectors, custodians, and consumers of data products. It should be borne in mind that the objective assessments could be task-independent, but could be task-dependent too. Task-independent metrics reflect states of the data without the contextual knowledge of the application and can be applied to any data set, regardless of the tasks at hand, while the task-dependent metrics are developed in specific application contexts [9]. According to our observations, most of the business significant data quality problems in modern organizations require the use of task-dependent measurement. Also, there is a common expectation that objective data quality improvement necessarily implies business value. Unfortunately, the limited awareness of what data quality improvement can truly imply often drives technical approaches that do not always translate into improving the business. Objective data quality metrics (such as a number of invalid values or percentage of missing data elements) may not necessarily be tied to the business's performance [10].

Many publications discuss various assessment metrics and indicators. Different types of indicators are discussed as appropriate for different quality dimensions. However, guidelines and good practices of how to relate and adapt the objective metrics to the business context and consumer expectations are rarely discussed. This trend is particularly pronounced in publications related to quality assurance in the Big data area. Thereby the impression that the quality of data and its measurement is an objective one is reinforced.

2.3. The Importance of Data Quality Requirements analysis

Another common mistake is to underestimate the preliminary analysis to identify the extent of data quality issues and to draw up the quality improvement strategy. The allocated time and resources for the analysis phase usually are not enough. Many business professionals are not aware that they have to become part of such analysis, so they are not interested enough in the project aims. Other participants in turn focus on analyzing and profiling the data quality state in the databases, instead of interviewing the business users about the problems encountered by them. The organization remains unable to see the big picture of the quality issues' impact over its business and it is not able to decide how to achieve maximum benefit with minimum effort. The methodologies for data quality improvement also reflect in the underestimation of the need of a holistic analysis. Almost all methodologies include activities that examine data schemes and perform interviews to reach a complete understanding of data, but only a few of them consider the data quality requirements analysis step, identifying data quality issues from the business perspective and collecting new target quality levels [11].

2.4. Whose Responsibility is to Provide Quality Data?

One of the main factors for the success of data quality initiatives is to understand correctly, whose responsibility it is to provide quality data. In our experience, the misconception that data quality is mainly an IT problem and it could be solved by purchasing sophisticated (and often expensive) data quality tools is very common. One of the reasons for such a delusion is the assumption that developers bear primary responsibility for ensuring that the software they write works properly, no matter which data is fed into it. Since they write the code, they have the power to control how an application will respond when it receives low quality data. In the real world, even the best programmers cannot foresee every type of data quality issue that could occur within their applications. Even if they did, the applications might end up being overloaded by functions that handle obscure data quality issues [12]. Actually, the organization and its processes are always more important than the technology since they are defined according to company strategy [13]. The business professionals must not only be involved, but also lead the data quality initiatives. This is a consequence of the contextual nature of the quality. In fact, the businesses know the context and the indented uses of data better than anyone does.

Given the strategic importance that data plays in competitive markets, high data quality can only be achieved if IT and business find a mutually intelligible language, a mode of operation, that overcomes the traditional disconnect that can be observed in so many corporations. Conceptually what needs to happen to enable business alignment is bringing together problem holders and problem owners. The problem holder is the person who experiences "pain" from a problem; a problem owner is the person who controls the resources needed to resolve it. The problems resolutions lie in making the problem owner feel some of the problem holder's "pain" so he becomes motivated to do something about the problem [14].

2.5. Systems Interoperation in the Modern Business

The lack of overall vision for the data flows between information systems of the organization and the lack of effective strategy for synchronizing the changes in them often lead to the presence of inconsistent, obsolete and contradictory data. The managers rely on the corporate data warehouses as a "single source of truth". However, in modern data warehouses, which accumulate a huge amount of data from different operating systems, this is difficult to achieve in practice. Moreover, the operating systems that feed data warehouses are also used for an extraction of various reports. Correction of bad data must be done in all systems where copies are stored, although the data may be used for different purposes in different systems. Correcting them only in one system (as part of an initiative for data quality improvement in a particular application) will improve the quality of data locally, but will enlarge another severe problem – the existence of conflicting information at the organizational level.

2.6. Identical Objects Recognition

A severe problem in some organizations is the resolution of cases with duplication of data, especially when there are no clear criteria by which it is possible to identify two objects as identical (e.g. if some individuals do not have the same national identification number in different data sources). Similar problems are particularly relevant when collecting information from various sources and information channels in large data warehouses, data lakes and Big data solutions. Duplicate information is a major source of inconsistency and lack of integrity in the organization's data. It is often associated with increased risks of submitting inaccurate information to external partners and regulatory authorities. The data integrity issues contribute greatly to the loss of trust in the organization's data, both by the company's employees and by customers, contractors and external partners, which in turn leads to a loss of reputation.

Ignoring or postponing the resolution of problems with duplicate objects is a common situation. To resolve such problems on a company level, organizations are looking for master data management solutions. Providing master data management across the enterprise is a complex task that requires investment of significant funds, allocation of a lot of human resources, amendments in already established and verified business processes, etc. However, the issues with duplicate objects can be solved (or at least minimized to a satisfactory level) with lightweight approaches. For example, efficient deduplication of objects can be achieved by using record linking techniques. In [3] and [15] we are sharing details of our implementation of a method for individuals and legal entities unification, which has been used productively for more than 15 years.

3. Challenges in Providing the Quality of Data Collected by WSNs

3.1. Quality of Data in Sensor Networks

With the emergence of the Internet of Things (IoT) and WSNs, sensor devices are deployed across the globe in a variety of fields such as healthcare, industry, agriculture, home, transport, etc. An IoT application may have hundreds or thousands of sensors, which produces vast amounts of data, but the data is useless if it is riddled with errors, as poor sensor data quality caused by the errors may lead to wrong decision-making results [16]. Ensuring the quality of monitoring data is fundamental to avoid false alarms or ignoring relevant data. However, because the sensors are located in the physical environment, they are constantly being subjected to factors that directly interfere with the data quality [17]. The hostile environment in which sensors are deployed plays a big part in the quality of the transmitted data [16]. For example, sensors for temperature, light, or humidity measurements are often placed outdoors and are subjected to potentially strong currents, debris accumulation and tough weather conditions, which might affect the operation of the sensors. Many errors in wireless sensor networks also occur due to characteristics, such as low-cost sensors, limited resources, and link variation. These errors appear in different modes, for example, the data loss or anomalies caused by hardware, the data failure due to transmission delays, and the sampling jitter caused by the node task conflicts. The dataset collected by the sink node (base station) may simultaneously result in all of these errors [18]. However, the WSN data could be incorrect not only due to a faulty sensor or corrupted communication. The erroneous data coming from sensors can be identified in two major categories: intentional or unintentional errors. Intentional data errors can be caused by physical or logical attacks over sensors (i.e. malicious attacks) and unintentional data errors are generally caused by hardware malfunction (or stochastic errors), misplacement (conditional errors) or exhausted resources (systematic errors) [19]. Consequently, there is a trust issue related to the collected data, which demands an extensive human intervention in terms of time and knowledge specialization, data validation tasks and periodic maintenance of sensors. To deal with this problem, it is necessary that the quality of collected data be continuously and automatically validated [17]. Simple strategies to overcome data quality problems, such as using industry grade sensors, which are more accurate, stable and robust, are not feasible for applications that require the deployment of large and dense sensors networks, which is the case for many IoT applications [16]. This approach will minimize a vast range of data errors, but it is far from solving all data quality issues and therefore it will not eliminate the problem with the lack of trust and the need of constant data validation. In addition, having to deploy many highly accurate but expensive sensors will incur higher deployment costs. It is therefore not surprising that in recent years automated software solutions to eliminate errors in data at the lowest possible architectural level have been actively developed. Intelligent sensors, self-calibrating and self-organizing protocols are emerging, which is an important necessity, especially if the WSN has to be located in hard-to-reach places.

In essence, data quality issues can be faced taking two main perspectives: (a) by deploying dependability techniques to mask faults and enforce the reliability of the system or (b) by enhancing the system with means to continuously assess and characterize the quality of data. In the first case, the system will not be aware of the quality of data, and hence, if a certain quality is needed, it must be enforced by design, confining the effects of faults a priori. Considering sensors to be the main source of data, errors in sensing measurements are handled by procedures that are established based on a deep understanding of the characteristics of the sensors. Missing readings may be handled by oversampling, and glitches, like outliers and noise, can be masked by averaging. In the second case, given that the system can be aware of the quality of data at run-time, it is better suited to be used in environments where full knowledge of the operational conditions is not known in advance. In this case, mitigation techniques must be deployed to handle faults and data quality problems at run-time, for instance exploiting application semantics to determine appropriate data corrections and to regain the needed data quality. Given that no system can be built to exhibit 100 % reliability, the two perspectives can be combined [17].

3.2. WSN Challenging Aspects

Many of the challenges in providing quality data collected from sensor networks are related to the following factors:

3.2.1. Environmental Conditions and Vandalism

As we mention above, the WSNs operations of sensor nodes are frequently susceptible to environmental effects. Wireless sensor networks allow surveillance of large areas that are remote, difficult to access, and/or dangerous to humans. Sensor devices are generally defenseless from outside physical threats, both from humans and from animals. Such acts often result in rendering sensors dysfunctional, which definitely affect the quality of produced data ([17], [20]). To monitor some phenomenon, sensors may be deployed in environments with extreme conditions (e.g. a mountain's summit). In such conditions, sensors may become dysfunctional or unstable due to many events (e.g. snow accumulation, dirt accumulation, etc.). Furthermore, the maintenance of such sensors is rarely ensured considering the inaccessibility of terrains [17]. In these scenarios, it is possible for multiple nodes to fail at the same time and partition the WSN into disjoint segments. Such loss of connectivity may cause service disruptions and render the WSN useless. Given the critical role a WSN plays and the fact that deployment of additional nodes may be infeasible, the WSN must have the ability to self-heal and restore connectivity by utilizing surviving resources [21]. Restoring network connectivity in a distributed, localized, efficient and intelligent on-demand manner is one of the major challenges in designing wireless sensor networks.

3.2.2. Low Cost Sensor Devices, Hardware and Resource Constraints

The things in the IoT suffer generally from a severe lack of resources (e.g. power, storage, etc.). Their computational and storage capabilities do not allow complex operations support. Furthermore, they are usually battery-powered and they often operate with discharged batteries [17]. Power is considered the most critical limitation among others as sensing and processing must be accomplished within a limited period of time. Computational and storage capabilities do not allow complex operations due, in turn, to the battery-power constraints among others. Resource management systems depend mostly on process migration by distributing the tasks among nodes [22]. For better network applicability, the nodes in the sensor networks are usually heterogeneous: they have different energies, functions, communication abilities, and so on. Some nodes in the network are endowed with stronger ability to enhance the network communication, which are called "super nodes" [23]. However, the ordinary sensors may lack precision or suffer from loss of calibration or even low accuracy, especially when they are of low cost. Faulty sensors may also result in inconsistencies in data sensing. A sensor node fails, but it keeps up reporting readings, which are erroneous. It is a common problem for sensor networks and an important source of outlier readings. Considering the scarce resources, data collection policies are adopted, which usually make trade-offs that affect the quality and cleanliness of data [17]. Since most traditional cleaning techniques have paid limited attention to reasonable availability of computational resources, a challenge for providing high data quality in WSNs is how to minimize the energy consumption while using a reasonable amount of memory for storage and computational tasks [24].

3.2.3. The Huge Volumes of Data and Architectural Complexity

While IoT sensors and devices collect tons of valuable insights, they also generate massive, high-speed data streams. As a point of reference, IoT devices and sensors can capture gigabytes of data within a few hours [25]. The huge volumes pose a number of challenges to the efficient transmission and processing of data, including in terms of methods and algorithms for quality assessment and data cleaning. The large number of devices in the sensor networks not only increases the amount of data, but also accumulates the chance of error occurrence. In addition, considerable amounts of sensed data are of no interest, meaningless, and redundant [26]. In such circumstance, it can be difficult to spot truly meaningful patterns. Therefore, a challenging aspect in sensory data processing is the need of a mechanism for extracting meaningful information from redundant data and providing a consistent, accurate and reliable data set in an efficient manner.

The management of the rapidly growing volumes is also a challenging task. As IoT adoption continues to rise, organizations from every sector struggle to keep up with these massive datasets expanding at exponential rates. This requires a robust architecture for efficient storage and retrieval of data. A high-level view of a typical IoT architecture includes three conceptual layers: Acquisition or Perception layer (measures and collects readings via sensor devices), Network layer (determines routes and transmits readings using wireless technologies such as WiFi, 2G/3G/4G, Bluetooth, LoRa, etc.), and Application layer. Typically, the Application layer is designed and implemented using Big data architectures such as Apache Hadoop, Spark, or Kafka. Virtually every layer has its own architectural challenges. The added complexity of the architecture causes new errors to be potentially introduced in each layer. Unlike typical wireless networks, WSNs have serious deficiencies in terms of data reliability and communication owing to the limited capabilities of the nodes. In the Network Layer, poor data quality arises from congested and unstable wireless communication links in sensor networks, which causes data loss and corruption. Omission, crash and delay faults as well as message corruption are the most common network errors that lead to DQ issues (data integrity issues, missing data, inaccurate data etc.) [27]. Design of a sensor network includes the hardware constraints, power constraints, operating environment, network architecture, routing topology etc. Finding a trade-off between the desired functionality and the available limitations when designing the network is a challenging task, but the expected accuracy of information is an important factor that is always considered [28]. In the Acquisition layer, damage to sensor devices or battery depletion are the cause of most data errors [16]. Typical sources of unreliable data at this architectural level are the so-called "soft faults" such as constant offset errors, drifts, stuck-at-zero errors, trimming errors, noises and outliers [16].

3.2.4. Data Redundancy

In Wireless Sensor Network (WSNs), nodes are densely deployed in a region to collect information. Sensors sense the similar data and forward to sink. This similar data sometimes leads to redundancy at the sink. The redundant data results in more accuracy, reliability and security whereas elimination helps in energy saving as most of the energy of a sink node gets wasted in dealing with the redundant data. Redundancy exploits to improve various data quality factors such as [29]:

• *Reliability*: Redundancy ensures the reliability in WSN. Redundant data produced by the sensors ensure more accurate data as compared by the single entry data. Sometimes sensor devices are providing imprecise measurements due to environment conditions. For this reason, reading from a single sensor may not provide reliable data. The reliability of the system could be enhanced by employing redundant data from the same sensor nodes or from other sensor nodes.

• *Accuracy*: If two or more independent sources collect the same piece of data over a period of time, these pieces can increase the associated confidence and provide accurate data.

• *Consistency*: Different data values bring conflicting information, this conflicting information is reduced by using spatial redundancy, temporal redundancy and information redundancy.

• *Detect malfunctioning or find out malicious nodes*: Analytical redundancy is used to detect malfunctioning or malicious sensors. This process uses a mathematical model in which past and present values of neighbor sensor nodes are used to predict the value of a sensor. This predicted value is then compared with real sensor value to find malicious sensors.

• *Completeness*: The redundancy allows implementation of techniques for online reconstruction of missing data.

Elimination of redundant data improves the lifetime of a sensor network whereas incorporation of redundancy enhances data quality. Data fusion or aggregation techniques used to deal with redundancy either drop the redundant data or use redundant data. The establishment of a mechanism that achieves both, i.e. reliable data as well as energy savings through utilizing redundancy is another challenging aspect in WSN [29].

3.2.5. Real-time Data Streaming

The need of real-time data processing in many sensor network applications presents a number of challenges. Some challenges are inherent to the nature of data streams, and the quality evaluation algorithm has to face these constraints. Classical cleaning and evaluation algorithm, which is designed to process static data, should be modified to face the real-time processing specifics. For example, in the case of real-time processing, data values arrive at a different rate, so the evaluation algorithm should provide a mechanism for processing existing data before the arrival of the new incoming data. Transient data is also a significant challenge. Data expires after a while and loses credibility, so data processing should be done before it expires. In addition, data distribution may change after some time, and if the evaluation algorithm fails to find the new distribution, it will not be adequately performed [30].

The machine learning algorithms have their special challenges in terms of real-time processing. For specificity, we will give some examples. Data preparation is an essential part of a machine learning solution. Real-world problems require transformations to raw data, preprocessing steps and usually a further selection of 'relevant' data before it can be used to build machine-learning models. Trivial preprocessing, such as normalizing a feature, can be complicated in a streaming setting. The main reason is that statistics about the data are unknown a priori, e.g., the minimum and maximum values a given feature can exhibit [31]. In most clustering algorithms, the number of clusters is one of the crucial parameters to be defined. Since there is no access to all data in the data stream, it is impossible to specify the number of clusters. Not specifying the number of clusters leads to the production of arbitrary shapes, and ultimately it will be difficult to analyze these clusters [30]. In the case of batch processing, the classification algorithms could produce a model with reasonable accuracy, which is trained and tested over a selected set of data, but in the streaming data, specifying the size of test data and updating the model are major challenges. Due to the challenging aspects of real-time data processing, so-called hybrid approaches have recently gained popularity. These approaches work both on batch and stream data, typically using static data to build a model and apply the model to the data stream.

It should be kept in mind that each particular business and even task may require a completely different approach to real-time analysis. Choosing the data collection and transmission interval that is right for the specific task purposes may also be challenging. Choosing the right interval so that it is not too frequent but allows valuable insights to be extracted in time is specific to every enterprise [32].

3.2.6. Data Security

WSNs and sensor nodes can be subject to attacks that may significantly affect the quality of sensor data, among other consequences for the application. Because the IoT delivers access to enterprise IT resources beyond an organization's protective firewalls, connected devices are often targeted by hackers. Endpoint devices in the IoT are vulnerable for many reasons. Since they are in external settings, endpoint devices do not have the same traditional perimeter security protection as centralized resources. IoT devices are designed to be small and lightweight in order to quickly emit data and they have the necessary hardware and processing power of infrastructure with more robust security (e.g. no support for cryptographic operations because of their high consumption of resources) [33]. In addition, it is difficult always to authenticate the devices in a network. These, as well as some other factors, lead to increased risks of attacks to alter data transmitted by a sensor device that can lead not only to wrong decisions, but also to triggering malicious actions or events (both by humans and by machines).

3.2.7. Establishing a Connection Between Errors in Data and Their Causes

The errors in WSN data can be classified in two basic categories: systematic errors and random errors. Generally speaking, the systematic errors could be compensated by self-tuning and self-calibration processes, while random errors should be removed or replaced by predicted values using anomaly detection techniques. However, data errors can have many different forms and causes. For example, errors that randomly occur in the data may be associated with a systematic cause (e.g. dust accumulated in the sensing unit). Establishing a link between errors and their causes is a challenging task. By having this link, the errors can be tolerated at their source [34].

This is one reason some of the anomaly detection methods to carry out further analysis to determine whether a real anomaly has been detected or it is a result of an event affecting certain segments of the network (e.g. a fire in a forest, a gas leakage in a coalmine, a flood in an agriculture field etc.). Virtually, the ability to determine the causes of data anomalies is a key point in selecting adequate error handling actions – automatic correction or error ignoring, triggering an alarm, notification of force majeure event, etc.

3.2.8. Self-calibration Complexity

The calibration procedures in uncontrolled environments could be a complex task. The sensor nodes are exposed to environmental noise and hardware failures, and the mismatch between factory calibration conditions and in-field conditions makes the calibration of sensors challenging. Moreover, different sensors require different calibration procedures, and the reference values might not be readily available. Therefore, in general, even when the response function is known, calibration in uncontrolled environments is difficult. An example is an ozone (O_3) sensor whose output value depends not only on the O_3 concentrations but also on the environmental conditions – temperature and relative humidity – of the place in which the sensor is deployed. Conducting a multiple linear regression with respect to ground-truth O_3 concentrations in a place with different environmental conditions than those of the place in which the sensor node will be deployed can produce large errors in the predicted O_3 concentrations [35]. The temperature sensors have a similar behavior – on-site calibrated sensors behave differently when placed in another location because of the difference in environmental conditions between the two locations (solar radiation, humidity, wind speed, rainfall, azimuth, etc.) [36].

3.3. In Summary

Sensor data comes very often with a variety of quality issues, posing many challenges. This can occur due to diverse reasons including power outages at sensors, network issues, sensor malfunctioning or external influences. Sensor data recovery is a great challenge due to the spatio-temporal nature and their stochastic distribution. Re-transmitting the data cannot be a meaningful solution in the context of IoT Analytics as most of the IoT applications operate in real-time data streams, which would mean that the delayed data is considered useless. Besides that, the computation and energy cost limit the sensor’s efficiency as these devices are usually constrained in terms of battery, memory, and computational resources. High-quality data input is imperative to empower to make quick data-driven decisions. Poor data including inaccurate, non-consistent, missing or incomplete data leads to inconvenient, time-consuming and expensive failures [37]. Although common types of errors produced by sensor networks have been well studied and a number of good solutions have been found to eliminate them, the issues with the completeness, accuracy, consistency and timeliness of data cannot be solved completely. Therefore, organizations working with WSN applications must identify the data quality issues and ensure the quality of sensor data streams in accordance with their goals and requirements. It is prudent that the established methodologies and proven practices be followed in providing quality data. Unfortunately, it is currently difficult to find DQ methodologies that are tailored to sensor networks. Numerous articles address individual aspects of providing quality data collected by sensor networks, but comprehensive DQ management methodologies are missing. In fact, we found only one examination of such methodology in [27]. However, we expect new studies in this area to appear in the near future.

4. Conclusions

Data quality has been a topic of interest to the scientific community for more than 30 years.

Accumulated knowledge and experience are significant, but there is still much to be done. The recent surveys show that companies are failing in their efforts to become data-driven. At present, many CEOs understand in principle the value and potential that data offers yet seem reticent to fully trust what the data tells them. KPMG International's recent Guardians of Trust study highlighted the extent of this conflict — just 35 percent of executives surveyed said they had a high level of trust in their own organization's use of data analytics while 25 percent admitted they either have limited levels of trust in, or actively distrust, the data they receive. In addition, 67 percent of CEOs say they have ignored the insights provided by data analysis or computer-driven models in recent years, because they have contradicted their own intuition or experience [38]. It is probably not a surprise, then, that 36 percent of CEOs admit they cannot make data-driven decisions until they invest significantly in data quality [39].

Our experience shows that when it comes to enterprise DQ initiatives most organizations have the same class of problems and obstacles to achieving better data quality. Many issues are a consequence of underestimating the importance of the business context and the domain knowledge, precise specification of the data use purposes and an active involvement of the domain experts. It is widely expected that IT and data science professionals are the people in charge of high-quality data providing, but in terms of data-driven decision-making, this is not a reasonable expectation. As far as the IoT and WSNs are concerned, the technical challenges that are in the hands of engineers and data science professionals are far more than in DQ provisioning for tactical business decisions. This is partly because the data quality issues in the sensor networks are focused on particular data uses. In general, the WSNs data are collected in a specific context for specific data uses and the professionals are looking for effective methods to clean up in this context. Therefore, it is not surprising that the discussed data quality issues in the field of IoT and WSNs have their unique challenging aspects, which are different from the challenges in data warehousing and strategic decisions. However, some data quality issues are in common. For instance, any DQ initiative, regardless of its type or context, must meet the needs of data consumers in order to define an effective data improvement strategy; the data quality metric, assessment and cleaning algorithms should be adapted to the business/domain context; the domain experts should be actively involved in the DQ initiative and requirements definition, etc.

In the current article, we have paid attention to only a few data quality issues and WSN challenges, which we consider as important enough to be addressed, but their list is much longer for open discussion. In our opinion, the efforts in the data quality area should be focused not only on the development of new techniques and methodologies, but also on the promotion of common mistakes, as well as the sharing of good practices.

Acknowledgement

This research is supported by the project BG05M2OP001-1.001-0004 Universities for Science, Informatics and Technologies in the e-Society (UNITe) and the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)".

References

[1]. N. Laranjeiro, S. Soydemir, J. Bernardino, A Survey on Data Quality: Classifying Poor Data, in *Proceedings of the IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC*', China, 18-20 November 2015, pp. 179-188.

[2]. F. Clemente, The cost of poor data quality, *Towards Data Science*, 2020 (https://towardsdatascience.com/the-cost-of-poor-data-quality-cd308722951f).

[3]. I. Naydenova, Z. Kovacheva, K. Kaloyanova, Important Data Quality Accents for Data Analytics and Decision Making, in *Proceedings of the 1st IFSA Winter Conference on Automation, Robotics & Communications for Industry 4.0 (ARCI'21)*, Chamonix-Mont-Blanc, France, 3-5 February 2021, pp. 90-95.

[4]. R. Wang, L. Guarascio, Dimensions of Data Quality: Toward Quality Data by Design, *IFSRC Discussion Paper #CIS-91-06* (https://web.mit.edu/smadnick/www/wp2/1991-06.pdf).

[5]. R. Wang, D. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, Vol. 12, Issue 4, 1996, pp. 5-34.

[6]. L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement - A Data Quality Assessment Framework, *Morgan Kaufmann*, 2013.

[7]. J. Todorov, I. Sturbanova, M. Trifonova, Information System for Planning, management and reporting of open Cast Mines Production (Output), in *Proceedings of the First International Conference on Information Systems & Datagrid*, Sofia, Bulgaria, 17-18 February, 2005, pp. 147-154, ISBN: 954-649-761-4.

[8]. ISO/IEC 25012:2008, Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model.

[9]. L. Pipino, Y. Lee, R. Wang, Data Quality Assessment, *Communications of the ACM*, Vol. 45, Issue 4, April 2002, pp. 211-218 (http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf).

[10]. D. Loshin, The Practitioner's Guide to Data Quality Improvement, *Morgan Kaufmann*, 2011.

[11]. C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for Data Quality Assessment and Improvement, *ACM Computing Surveys*, Vol. 41, Issue 3, 2009, Article 16.

[12]. Ch. Tozzi, Solving Data Quality Problems Is Not (Only) Programmers' Responsibility, 2020, Precisely (https://www.precisely.com/blog/data-quality/solving-data-quality-problems).

[13]. BARC, Data Quality and Master Data Management: How to Improve Your Data Quality, 2020, *Business Application Research Center*

(https://bi-survey.com/data-quality-master-data-management).

[14]. T. Breur, Data quality is everyone's business - Managing information quality - Part 2, *Journal of Direct, Data and Digital Marketing Practice,* Vol. 11, 2009, pp.114–123 (https://link.springer.com/article/10.1057/dddmp.2009.21).

[15]. I. Naydenova, K. Kaloyanova, S. Ivanov, Multi-Source Customer Identification, in *Proceedings of the Third International Conference on Information Systems & Grid Technologies*, 28 - 29 May 2009, Sofia, Bulgaria, pp. 77-85.

[16]. H. Teh, A. Kempa-Liehr, K. Wang, Sensor data quality: a systematic review, *Journal of Big Data*, Vol. 7, 2020, Article number: 11 (https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1).

[17]. G. Jesus, A. Casimiro, A. Oliveira, A Survey on Data Quality for Dependable Monitoring in Wireless Sensor Networks, *Sensors,* 2017, Vol. 17, Issue 9, 2010 (https://www.mdpi.com/1424-8220/17/9/2010/htm).

[18]. H. Cheng, D. Feng, X. Shi, Ch. Chen, Data quality analysis and cleaning strategy for wireless sensor networks, *EURASIP Journal on Wireless Communications and Networking,* 2018 (https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-018-1069-6).

[19]. C. Rodriguez, S. Servigne, Managing Sensor Data Uncertainty: a data quality approach, *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, Vol. 4, Issue 1, 2013, pp. 35-54 (https://hal.archives-ouvertes.fr/hal-01339140/document).

[20]. R. Perez-Castillo, A. Carretero, M. Rodriguez, I. Caballero, M. Piattini, Data Quality Best Practices in IoT Environments, in *Proceedings of the 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, Coimbra, Portugal, 2018, pp. 272-275 (https://intelligentenvironments.es/images/files/QUATIC2018_SHORT.pdf).

[21]. Y. Joshi, M. Younis, Restoring connectivity in a resource constrained WSN, *Journal of Network and Computer Applications*, Vol. 66, 2016, pp. 151-165 (https://www.sciencedirect.com/science/article/abs/pii/S1084804516300212).

[22]. L. Tawalbeh, S. Hashish, H. Tawalbeh, Quality of Service requirements and Challenges in Generic WSN Infrastructures, *Procedia Computer Science*, Vol. 109, 2017, pp. 1116-1121.

[23]. Z. Zhang, J. Li, X. Yang, Data Aggregation in Heterogeneous Wireless Sensor Networks by Using Local Tree Reconstruction Algorithm, *Complexity*, Vol. 2020, 2020, pp. 1-14.

[24]. Y. Zhang, N. Meratnia, P. Havinga, Outlier Detection Techniques for Wireless Sensor Networks: A Survey, *IEEE Communications Surveys & Tutorials*, Vol. 12, Issue 2, pp. 159-170.

[25]. Real-Time Processing of Data for IoT Applications, *Tiempo Development,* 2020, Blog Post (https://www.tiempodev.com/blog/real-time-processing-of-data-for-iot-applications/).

[26]. B.-S. Kim, K.-I. Kim, B. Shah, F. Chow, K. H. Kim, Wireless Sensor Networks for Big Data Systems, *Sensors (Basel)*, Vol. 19, Issue 7, 2019, 1565. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6480280/).

[27]. R. Perez-Castillo, A. Carretero, I. Caballero, M. Rodriguez, M. Piattini, A. Mate, S. Kim, D. Lee, DAQUA-MASS: An ISO 8000-61 Based Data Quality Management Methodology for Sensor Data, *Sensors (Basel)*, Vol. 18, Issue 9, 2018, 3105. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6165119/)

[28]. S. Halde, K. Khot, Big Data in Wireless Sensor Network: Issues & Challenges, *International Journal of Advanced Engineering, Management and Science*, Vol. 2, Issue 9, 2016, pp. 1618-1621.

[29]. N. Verma, D. Singh, Data Redundancy Implications in Wireless Sensor Networks, *Procedia Computer Science*, Vol. 132, 2018, pp. 1210-1217. (https://www.sciencedirect.com/science/article/pii/S1877050918307683).

[30]. M. Mirzaie, B. Behkamal, S. Paydar, State of the Art on the Quality of Big Data: A Systematic Literature Review and Classification Framework, *Computer Science,* 2019. (https://arxiv.org/vc/arxiv/papers/1904/1904.05353v1.pdf).

[31]. H. Gomes, J. Read, A. Bifet, J. Barddal, J. Gama, Machine learning for streaming data: state of the art, challenges, and opportunities, *ACM SIGKDD Explorations Newsletter*, Vol. 21, Issue 2, November 2019, pp. 6-22.

[32]. Imranali, 5 major sensor data analytics challenges: deadly or curable?, *Data Science Central Blog Post*, *TechTarget, Inc*, 2018 (https://www.datasciencecentral.com/profiles/blogs/5-major-sensor-data-analytics-challenges-deadly-or-curable).

[33]. IoT Security: What it is and Best Practices, *Talend Web Portal*, 2021. (https://www.talend.com/resources/what-is-iot-security/).

[34]. X. Fnag, Improving data quality for low-cost environmental sensors, *PhD Thesis, University of York Computer Science*, 2018 (https://etheses.whiterose.ac.uk/21259/7/XINWIFANG_PHD_Thesis.pdf).

[35]. J. Barcelo-Ordinas, M. Doudou, J. Garcia-Vidal, N. Badache, Self-Calibration Methods for Uncontrolled Environments in Sensor Networks: A Reference Survey, *Ad Hoc Networks*, *Elsevier*, Vol. 88, 2019, pp. 142-159 (https://hal.archives-ouvertes.fr/hal-02162764/document).

[36]. K. Yamamoto, T. Togami, N. Yamaguchi, S. Ninomiya, Machine learning-based calibration of low-cost air temperature sensors using environmental data, *Sensors,* Vol. 17, Issue 6, 2017, 1290.

[37]. A. Joseph, A. Sharma, IoT Analytics: Data Quality challenges, *Tech Mahindra*, 2020 (https://cache.techmahindra.com/static/img/pdf/iot-analytics-pov-modified-3Aug2020.pdf)

[38]. B. Fisher, Five ways CEOs can build a data-driven organization, *Morgan Kaufmann*, 2019, KPMG (https://home.kpmg/xx/en/home/insights/2019/04/dont-doubt-the-data.html)

[39]. KPMG, Trusted analytics matter to CEOs, *Trusted Analytics Article Series*, Issue 7, August 2017. (https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/08/da-trusted-analytics-ceo-outlook.pdf).