Data Quality Best Practices in IoT Environments

Ricardo Perez-Castillo[1], Ana G. Carretero[1], Moises Rodriguez[1,2], Ismael Caballero[1], Mario Piattini[1]
[1] Information Technologies & Systems Institute, University of Castilla-La Mancha
[2] AQC Lab
Camino de los Moledores s/n 13051, Ciudad Real, Spain
[ricardo.pdelcastillo | anaisabel.gomez | moises.rodriguez | ismael.caballero | mario.piattini]@uclm.es

Alejandro Mate
Lucentia Lab
University of Alicante
San Vicente s/n. 03690. Alicante, Spain
amate@lucentialab.es

Sunho Kim
Department of Industrial & Management Engineering,
Myongji University, Korea
shk@mju.ac.kr

Dongwoo Lee
GTOne
501, 2Dong, Ace Hitechcity Bldg., 775,
Gyeongin-ro, Yeongdeungpo-gu, Seoul,Korea
leewow@gtone.co.kr

***Abstract*—The Internet-of-Things (IoT) is a network of physical devices embedded with electronics, sensors, actuators, and connectivity enabling these objects to connect and exchange data. The IoT brings more data than ever before, connecting cyber and physical worlds, which enable people to make better decisions. The very nature of the IoT environment generating giant volumes of data gathered from several heterogeneous sources creates important data quality challenges that need to be addressed. This is because inadequate levels of data quality impact negatively all the working of the IoT network. Despite the fact that data quality has been broadly studied outside the IoT field, and huge standardization efforts are behind the concept, data quality management in the IoT has not been investigated in-depth. This paper presents an approach for managing data quality in 'smart, connected product (SCP)' environments, grounded on the series of international standards ISO/IEC 25000 and ISO 8000. Our approach provides a set of best practices for assessing and improving data quality in such environments.**

Keywords—Data Quality; Internet-of-Things; Industry 4.0;

I. INTRODUCTION

"Our economy, society and survival aren't based on ideas or information—they're based on things" [1]. This is one of the core foundations of the Internet-of-Things (IoT) as stated by K. Ashton, who coined the term. IoT is an emerging global Internet-based information architecture facilitating the exchange of goods and services [2]. Thus, IoT systems are inherently built on data gathered from heterogeneous sources in which the volume and speed of data generation are dramatically increasing [3]. Furthermore, there is a certain emergence of IoT semantic-oriented vision which needs ways to represent and manipulate the huge amount of raw data expected to be generated from the "things"[4]. The vast amount of data in the IoT environments, gathered from a global-scale deployment of smart-things, is the basis for making intelligent decisions and providing better services. In other words, data represents the bridge that connects cyber and physical worlds. Despite of its tremendous relevance, if data are of poor quality, decisions are likely to be unsound [5, 6]. As a result, Data Quality (DQ) has become one of the key aspects in IoT [5, 7-9]. The IoT, and in particular smart, connected products (SCP), has concrete characteristics that favour the aparition of problems due to inadequate levels of data quality. While some of these characteristics might be considered omnipresent (i.e. uncertain, erroneous, noisy, distributed and voluminous), other characteristics are more specificand highly dependant on the context and monitored phenomena (i.e. smooth variation, continuous, correlation, periodicity or Markovian behaviour) [5].Outside of the IoT research area, DQ has been broadly studied during last years, and it has become into one of the more mature research area capturing the growing interest of the industry. This fact is reflected by the standardization efforts like ISO/IEC 25000 series addressing systems and software quality requirements and evaluation (SQuaRE) [10], or ISO 8000-60 series concerning the best practices in data quality management processes. In order to assure adequate levels of data quality, it is necessary to produce and implement methods, processes, and specific techniques for managing data concerns.Due to the youth of IoT, and despite DQ standards, frameworks, management techniques and tools proposed in the literature, DQ for IoT has been mostly ignored. Some works have addressed some DQ concerns in sensor wireless networks [7, 11], or in data streaming [12, 13] among other proposals [5]. Unfortunately, none of these works have considered the management of DQ in a holistic way in line with existing DQ-related standards. Thereby, this paper provides practitioners and researchers with some DQ best practices for improving quality of data in SCP environments grounded on / aligned to ISO 8000-62 [14]

The remainder of this paper is organized as follows: Section II lists some challenges concerning Data Quality in SCP environments. Section III presents our framework for Data Quality management in IoT. Finally, Section IV provides conclusions of this work.

II. DATA QUALITY CHALLENGES IN SCP ENVIRONMENTS

According to Cook et al. [15], a Smart Environment is a small world where all kinds of smart devices are continuously working to make inhabitants' lives more comfortable. Mühlhäuser [16] defines smart, connected products (SCP) as *"entities (tangible object, software, or service) designed and*

978-1-5386-5841-3/18/$31.00 ©2018 IEEE
DOI 10.1109/QUATIC.2018.00048

made for self-organized embedding into different (smart) environments in the course of its lifecycle, providing improved simplicity and openness through improved connections". SCP have three core components: physical, smart, and connectivity components. Smart components extend the capabilities and value of the physical components, while connectivity extends the capabilities and value of the smart components. This enables some smart components to exist outside the physical product itself, with a virtuous cycle of value improvement [17].

SCP can be connected in large, complex networks throughout three different layers [8]. **Acquisition layer** refers to the sensor data collection system where sensors, raw (or sensed) and pre-processed data are managed. **Processing layer** involves data resulting from data processing and management centre where energy, storage and analyse capabilities are more significant. **Utilization layer** concerns delivered data (or post-processed data) exploited.

Networking and management of SCP operations can generate the business intelligence needed to deliver smart services. Smart services are delivered to or via intelligent objects that feature awareness and connectivity [18]. According to these characteristics, SCP operations enble new capabilities for companies, arising new problems and challenges that must be taken into account. There are some SCP characteristics that can threat adequate data quality levels in data coming from SCP networks [5]. Such factors and related problems often result in dysfunctional SCP devices and sensors which definitely affect the quality of produced data (see TABLE I.). The dysfunctional SCP devices are the cause and having data with inadequate levels of quality is the consequence. Therefore, while we can identify dysfunctional SCP devices through the generation of data with inadequate data quality levels, it is noteworthy that these devices will impact the rest of the IoT network reading poor quality data generated by them. This section provides the challenges related to DQ management in SCP operations. First, SCP characteristics that affect DQ are presented. Then, all concrete problems due to inadequate levels of DQ in SCP environments are detailed. SCP characteristics affecting DQ.

A. DQ Issues in SCP

Tilak et al. [19] provide a taxonomy of sensor errors. Such errors are directly related to different data quality problems in the acquisition layer. The mentioned taxonomy distinguishes the following six types of data sensors errors (see TABLE II.). Apart of errors in isolated SCP devices, there are other communication errors which can happen at SCP network level [19] such as omission, crashes, delay and message corruption. All the mentioned SCP devices / sensor errors lead to different DQ problems in the three layers comented before. As said, DQ problems can be represented by means of some DQ dimensions or characteristics that are specially affected in different environemnets.

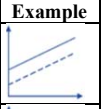
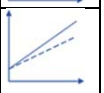
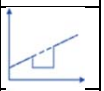
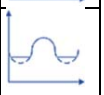
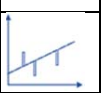
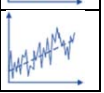
III. DATA QUALITY MANAGEMENT IN IOT

Reviewing the literature, DQ has been defined in different ways. The widest used definitions are aligned with the concept of "fitness for use". In [20], it is defined as: "Data Quality is data that is fit for use by data consumer. This means that usefulness and usability are important aspects of quality". Different stakeholders can have different perceptions of what quality means for the same data [21]. It largely depends on the context in which data is to be used. Thus, DQ in IoT environments must be managed in a specific way. ISO/IEC 25012 [22] defines a Data Quality Model widely used, which focuses on the quality of the data as part of an information system and defines quality characteristics for target data used by humans and.

TABLE I. SCP CHARACTERISTICS AFFECTING DQ (ADAPTED FROM [5])

SCP factor	DQ effect
Deployment Scale	IoT is expected to be deployed on a global scale. This leads to an huge heterogeneity in data sources (not only computers but also daily objects). Also, the huge number of devices accumulates the chance of error occurrence.
Resources constraints	Like for example computational and storage capabilities that do not allow complex operations due, in turn, to the battery-power constraints among others.
Network	Intermittent loss of connection in the IoT is recurrent. Things are only capable of transmitting small-sized messages due to their scarce resources.
Sensors	Embedded sensors may lack precision or suffer from loss of calibration or even low accuracy. Faulty sensors may also result in inconsistencies in data sensing.
Environment	SCP devices will not be deployed only in tolerant and less aggressive environments. To monitor some phenomenon, sensors may be deployed in environments with extreme conditions. Data errors emerge when the sensor experiences the surrounding environment influences [19].
Vandalism	Things are generally defenseless from outside physical threats (both from humans and animals).
Fail-dirty.	A sensor node fails, but it keeps up reporting readings which are erroneous. It is a common problem for SCP networks and an important source of outlier readings.
Privacy	Privacy preservation processing, thus DQ could be intentionally reduced.
Security vulnerability	Sensor devices are vulnerable to attack, e.g., it is possible for a malicious entity to alter data in an SCP device.
Data stream processing.	Data gathered by smart things are sent in the form of streams to the back-end pervasive applications which make use of them. Some stream processing operators could affect quality of the underlying data [9].

TABLE II. SENSORS ERRORS DERIVING DQ PROBLEMS IN SCP ENVIRONMENTS (ADAPTED FROM [19]).

Error	Description	Example
Constant or offset error	The observations continuously deviate from the expected value by a constant offset.	
Continuous varying or drifting error	The deviation between the observations and the expected value is continuously changing according to some continuous time-dependent function (linear or non-linear).	
Crash or jammed error	The sensor stops providing any readings on its interface or gets jammed and stuck in some incorrect value.	
Trimming error	Data is correct for values within some interval, but are modified for values outside the interval. Beyond the interval, the data can be trimmed or may vary proportionally.	
Outliers error	The observations occasionally deviate from the expected value, at random points in the time domain.	
Noise error	The observations deviate from the expected value stochastically in the value domain and permanently in the temporal domain.	

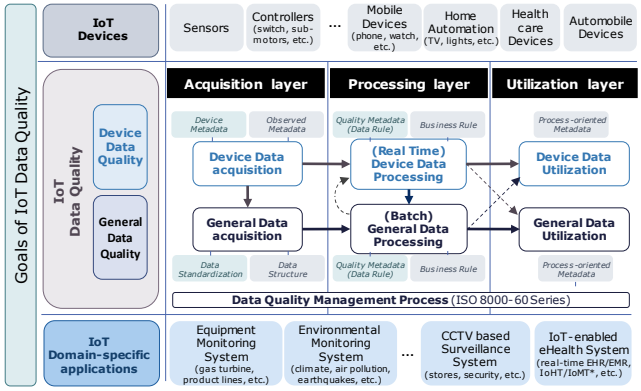


Fig. 1. IoT Data Quality conceptual framework

TABLE III. MATURITY MODEL FOR DQ MANAGEMENT IN IOT

Level	Id.	Process	Best practice
1	DQC.2	Data Processing	BP1, BP2
	DRS.4	Data Security Management	BP3, BP4
2	DQP.1	Requiremients Management	BP6
	DQC.1	Provision of Data Specifications and Work Instructions.	BP7
	DQC.3	Data Quality Monitoring and Control.	BP8,BP9,BP10, BP11, BP12
3	DQP.2	Data Quality Strategy Management.	BP14, BP23
	DQP.3	Data Quality Policy/ Standards/Procedures Management.	BP5,BP15, BP16, BP17
	DQP.4	Data Quality Implementation Planning.	BP13
	DRS.1	Data Architecture Management.	BP18, BP19
	DRS.3	Data Operations Management.	BP20, BP21
	RPV.1	Data Quality Organization Management.	
4	DQA.1	Review of Data Quality Issues.	
	DQA.2	Provision of Measurement Criteria.	
	DQA.3	Measurement of Data Quality and Process Performance.	
	DQA.4	Evaluation of Measurement Results.	
	DRS.2	Data Transfer Management.	
	RPV.2	Human Resource Management.	BP22
5	DQI.1	Root Cause Analysis and Solution Development.	
	DQI.2	Data Cleansing.	
	DQI.3	Process Improvement for Data Nonconformity Prevention.	

ISO/IEC 25012 provides a data-product centric vision. However, it still possible to consider a process-centric vision which identify data related processes (data management, data quality management and data governance). In this sense, the family ISO/IEC 8000-6x [23] provides a set of components that can help organizations to guide the assessment and improvement of their maturity when running such data quality management processes. While 8000-60 provides a global vision of the parts for data quality management; 8000-61 provides a processes reference model of data quality management processes and 8000-62 provides an evaluation model of data quality management processes maturity.

This section presents our main contribution: a conceptual framework for holistic data quality management in IoT environments. First, Fig. 1 shows on overview of the conceptual IoT Data Quality Management framework proposed. This general framework is extensible and can consider different kinds of IoT devices (top part of Fig. 1) and certain domain-specific applications (bottom part of Fig. 1).

The conceptual framework is ISO 8000-60 series [23] aligned. The DQIoT framework also provides methods to measure, manage, and improve data quality. The conceptual framework concerns are then considered throughout the three layers previously explained: acquisition, processing and utilization. All the three layers consider both; device-aware DQ and the quality of general-purpose data. This paper focuses on the sensor-aware DQ. Thus, this section focuses on the SCP devices / sensor DQ by listing some good practices aligned to ISO 8000-61 to enable organizations to be more efficient when executing DQ management activities in IoT environments.

Some best practices considered are based on the issues considered in [24] for quality assurance and quality control of sensor data. Other best practices are incorporated from our conclusions made during our investigations. The following list enumerates (BP1 to BP23) and describes 23 best practices. The implementation of these best practices for DQ management can lead to a improved DQ management and can also these affect to the measurement of the DQ characteristics.

BP1: Ensure that the sensor data are collected sequentially.

BP2: Document all sensor data processing.

BP3: Address security throughout the device lifecycle, from the initial design to the operational environment [25].

BP4: Identify risk for SCP operations.

BP5: Use flags to convey information about the data.

BP6: Record the date and time of known events that may affect measurements.

BP7: Provide complete metadata.

BP8: Maintain an appropriate level of human inspection.

BP9: Implement an automated alert system to warn about potential sensor network issues.

BP10: Perform range checks on numerical data.

BP11: Perform domain checks on categorical data.

BP12: Perform slope and persistence checks on continuous data streams.

BP13: Make available ready access to replacement parts.

BP14: Schedule sensor maintenance and repairs to minimize data loss.

BP15: Automating sensor data quality procedures.

BP16: Document all quality procedures that were applied

BP17: Create policies for comparing data with data from related sensors.

BP18: Create a strategy for node replacement [26, 27].

BP19: Replicating data sensors.

BP20: Retain the original unmanipulated data.

BP21: Retain all versions of the input data, workflows, data provenance programs, and models used (data provenance).

BP22: Resources management

BP23: Lifecycle management

IV. CONCLUSIONS

This paper dealt with the challeging problems of considering data quality in the IoT. Many research and standarization efforts have been made in the DQ area over the last years, and some interesing results have been already transferred to IoT as well. However, the approaches presented in the literature have two main drawbacks. On the one hand, as collected in this paper, many concrete factors associated with the nature of SCP, which have not been considered by existing proposals, and in general with IoT, affect the way in which DQ can/must be treated in such context. On the other hand, many DQ management standards have not been still succesfully tailored for the IoT case , and more specifically to SCP.

As a main contribution, we have consequently provided a set of tailored best practices for improving DQ management in SCP environments. In addition, we analysed such practices and the aligned them with ISO/IEC 25012 characteristics. A second contribution consisted of a preliminary maturity model for IoT based on ISO 8000-62 model including the best practices identificated for the processes of part 8000-61 at each maturity level. The main implication for researchers and practitioners is that they can consider these best practices as a mean to ensure DQ or to establish DQ controls in SCP or IoT environments. Furtehrmore, these best practices can be carried out together with well-depicted processes in international standards.

The future work will consist in conducting a formal gap analysis for ISO/IEC 8000-61 and -62 to discover how they could be tailored to meet the specifics of the SCP operations, in order to propose a formal and well depicted specialization of such standards for IoT environments.

ACKNOWLEDGMENTS

This work was primarily supported by DQIoT project (Eureka programme, E!11737; and CDTI (Centro Para el Desarrollo Tecnológico Industrial), INNO-20171086). Additionally, this work was partially funded by SEQUOIA project (TIN2015-63502-C3-1-R) (MINECO/FEDER); ECD project (Evaluación y Certificación de la Calidad de Datos) (PTQ-16-08504) (Torres Quevedo Programme, MINECO). Finally, it was also supported through a grant Dr. Ricardo Pérez-Castillo enjoys from JCCM within the RIS3 initiatives.

REFERENCES

[1] Ashton, K., *That 'internet of things' thing.* RFID journal, 2009. **22**(7): p. 97-114.

[2] Weber, R.H., *Internet of things–governance quo vadis?* Computer Law & Security Review, 2013. **29**(4): p. 341-347.

[3] Hassanein, H.S. and S.M. Oteafy. *Big Sensed Data Challenges in the Internet of Things*. in *Distributed Computing in Sensor Systems (DCOSS), 2017 13th International Conference on*. 2017. IEEE.

[4] Atzori, L., A. Iera, and G. Morabito, *The internet of things: A survey.* Computer networks, 2010. **54**(15): p. 2787-2805.

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

[6] Merino, J., I. Caballero, B. Rivas, M. Serrano, and M. Piattini, *A data quality in use model for big data.* Future Generation Computer Systems, 2016. **63**: p. 123-130.

[7] Jesus, G., A. Casimiro, and A. Oliveira, *A Survey on Data Quality for Dependable Monitoring in Wireless Sensor Networks.* Sensors, 2017. **17**(9): p. 2010.

[8] Gutiérrez Rodríguez, C.C. and S. Servigne, *Managing Sensor Data Uncertainty: A Data Quality Approach.* International Journal of Agricultural and Environmental Information Systems (IJAEIS), 2013. **4**(1): p. 35-54.

[9] Klein, A., G. Hackenbroich, and W. Lehner. *How to Screen a Data Stream-Quality-Driven Load Shedding in Sensor Data Streams*. in *ICIQ*. 2009.

[10] ISO/IEC, *ISO/IEC 25000:2014. Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE*. 2014.

[11] Qin, Z., Q. Han, S. Mehrotra, and N. Venkatasubramanian, *Quality-aware sensor data management*, in *The Art of Wireless Sensor Networks*. 2014, Springer. p. 429-464.

[12] Campbell, J.L., L.E. Rustad, J.H. Porter, J.R. Taylor, E.W. Dereszynski, J.B. Shanley, C. Gries, D.L. Henshaw, M.E. Martin, and W.M. Sheldon, *Quantity is nothing without quality: Automated QA/QC for streaming environmental sensor data.* BioScience, 2013. **63**(7): p. 574-585.

[13] Klein, A. and W. Lehner, *Representing data quality in sensor data streaming environments.* Journal of Data and Information Quality (JDIQ), 2009. **1**(2): p. 10.

[14] ISO, *ISO/DIS 8000-62: Data quality -- Part 62: Data quality management: Organizational process maturity assessment: Application of the ISO/IEC 330xx family of standards*. 2018.

[15] Cook, D. and S.K. Das, *Smart environments: Technology, protocols and applications*. Vol. 43. 2004: John Wiley & Sons.

[16] Mühlhäuser, M., *Smart products: An introduction*, in *European Conference on Ambient Intelligence*. 2007, Springer. p. 158-164.

[17] Porter, M.E. and J.E. Heppelmann, *How smart, connected products are transforming competition.* Harvard Business Review, 2014. **92**(11): p. 64-88.

[18] Wuenderlich, N.V., K. Heinonen, A.L. Ostrom, L. Patricio, R. Sousa, C. Voss, and J.G. Lemmink, *"Futurizing" smart service: implications for service researchers and managers.* Journal of Services Marketing, 2015. **29**(6/7): p. 442-447.

[19] Tilak, S., N.B. Abu-Ghazaleh, and W. Heinzelman, *A taxonomy of wireless micro-sensor network models.* ACM SIGMOBILE Mobile Computing and Communications Review, 2002. **6**(2): p. 28-36.

[20] Strong, D.M., Y.W. Lee, and R.Y. Wang, *Data quality in context.* Communications of the ACM, 1997. **40**(5): p. 103-110.

[21] Wang, R.Y., *A product perspective on total data quality management.* Communications of the ACM, 1998. **41**(2): p. 58-65.

[22] ISO/IEC, *ISO/IEC 25012: Software engineering-software product quality requierements and evaluation (SQuaRE) - data quality model*. 2008.

[23] ISO, *ISO/TS 8000-60: Data quality -- Part 60: Data quality management: Overview*. 2017.

[24] Campbell, J.L., L.E. Rustad, J.H. Porter, J.R. Taylor, E.W. Dereszynski, J.B. Shanley, C. Gries, D.L. Henshaw, M.E. Martin, W.M. Sheldon, and E.R. Boose, *Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data.* BioScience, 2013. **63**(7): p. 574-585.

[25] Riahi, A., Y. Challal, E. Natalizio, Z. Chtourou, and A. Bouabdallah. *A systemic approach for IoT security*. in *Distributed Computing in Sensor Systems (DCOSS), 2013 IEEE International Conference on*. 2013. IEEE.

[26] Dhillon, S.S. and K. Chakrabarty. *Sensor placement for effective coverage and surveillance in distributed sensor networks*. in *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*. 2003. IEEE.

[27] Pan, J., L. Cai, Y.T. Hou, Y. Shi, and S.X. Shen, *Optimal Base-Station Locations in Two-Tiered Wireless Sensor Networks.* IEEE Transactions on Mobile Computing, 2005. **4**(5): p. 458-473.