Data Quality and Security Evaluation Tool for Nanoscale Sensors

Leon Reznik
Department of Computer Science
Rochester Institute of Technology
Rochester, New York, USA
e-mail: lr@cs.rit.edu

Sergey Edward Lyshevski
Department of Electrical and Microelectronic Engineering
Rochester Institute of Technology
Rochester, New York, USA
e-mail: Sergey.Lyshevski@mail.rit.edu

Abstract— **The paper proposes a novel approach to data and information management in multi-stream data collection systems with heterogeneous data sources. Data may be produced by novel nanoscale photonic, optoelectronic and electronic devices. Poor quality characteristics are expected. In the proposed approach, we use a set of data quality indicators with each data entity, and, develop the calculus that integrates various data quality (DQ) indicators ranging from traditional data accuracy metrics to network security and business performance measures. The integral indicator will calculate the DQ characteristics at the point of data use instead of conventional point of origin. The DQ metrics composition and calculus are discussed. The tools are developed to automate the metrics selection and calculus procedures for the DQ integration is presented. The user-friendly interactive capabilities are illustrated.**

Keywords **- data quality; computer security evaluation; data accuracy; data fusion**.

I. INTRODUCTION

Recently we entered a new era of an exponential growth of data collected and made available for various applications. The existing technologies are not able to handle such big amounts of data. This phenomenon was called the big data. Photonics and nanotechnology enabled microsystems perform multiple generations and fusions of multiple data streams with various data quality [1-6]. The development and application of quantum-mechanical nanoscale electronic, photonic, photoelectronic communication, sensing and processing devices significantly increase an amount of data which can be measured and stored. These organic, inorganic and hybrid nanosensors operate on a few photons, electrons and photon-electron interactions [1, 2, 4, 6]. Very low current and voltage, high noise, large electromagnetic interference, perturbations, dynamic non-uniformity and other adverse features result in heterogeneous data with high uncertainty and poor quality. The super-large-density quantum and quantum-effect electronic, optoelectronic and photonic nanodevices and waveguides are characterized by: (i) Extremely high device switching frequency and data bandwidth (~1000 THz); (ii) Superior channel capacity (~10^{13} bits); (iii) Low switching energy (~10^{-17} J) [7, 8]. The importance of DQ analysis, data enhancements and optimization is emphasized due to: (1) Low signal-to-noise ratio (ratio of mean to standard deviation of measured signals is ~0.25 in the emerged electrons-photons interaction devices); (2) High probability of errors (p is ~0.001); (3) High distortion measure, reaching ~0.1 to 0.3; (4) Dynamic response and characteristic non-uniformity. These characteristics must be measured, processed and evaluated and provided to a data used along with the data.

New generations of information systems provide communication and networking capabilities to transfer, fuse, process and store data. Various applications require the data delivery from their origin to the point of use that might be far away. The data transfer may lead to information losses, attenuation, distortions, errors, malicious alterations, etc. Security, privacy and safety aspects of data communication and processing systems nowadays play a major role and may have a dramatic effect on the quality of data delivered.

New DQ management methods, quality evaluation and assurance (QE/QA) tools and robust algorithms are needed to ensure security, safety, robustness and effectiveness. As the amount of data available multiplies every year, current information systems are not capable to process these large data arrays to make the best decision. Big data applications require better data selection of high quality inputs. The absence of DQ indicators provided along with the data hinders the recognition of potential calamities and makes data fusion and mining procedures as well as decision making prone to errors.

In this paper we offer a novel approach to the data management in information systems. We propose to associate the DQ indicators with each data entity, and, replace one-dimensional data processing and delivery with multi-dimensional data processing and delivery along with the corresponding DQ indicators. To realize this approach, we describe the structure and content of these DQ indicators, develop the calculus of processing, and, develop interactive tools to automate this process. The current situation in DQ research is described in Section II. The DQ metrics composition is presented in Section III, while the DQ calculus is reported in Section IV. The CAD tools are documented in Section V. The conclusions are outlined in Section VI.

 ISBN: 978-1-61208-376-6

II. CURRENT ENVIRONMENT AND ACHIVEMENTS IN DQ EVALUATION

DQ represents an open multidisciplinary research problem, involving advancements in computer science, engineering and information technologies. The studied problems are directly applicable in various applications. It is essential to develop technologies and methods to manage, ensure and enhance quality of data. Related research in a networking field attempts to investigate how the network characteristics, standards and protocols can affect the quality of data collected and communicated through networks. In sensor networks, researchers started to investigate how to incorporate DQ characteristics into sensor-originated data [9]. Guha *et al.* proposed a single-pass algorithm for high-quality clustering of streaming data and provided the corresponding empirical evidence [10]. Bertino *et al.* investigated approaches to assure data trustworthiness in sensor networks based on the game theory [11] and provenance [12]. Chobsri *et al.* examined the transport capacity of a dense wireless sensor network and the compressibility of data [13]. Dong and Yinfeng attempted to optimize the quality of collected data in relation to resource consumption [14],[15]. Current developments are based on fusing multiple data sources with various quality and creating big data collections. Novel solutions and technologies, such as nano-engineering and technology are emerged in order to enable DQ assessment. Reznik and Lyshevski outlined integration of various DQ indicators representing different schemes ranging from measurement accuracy to security and safety [16], as well as micro- and nano engineering [17]. The aforementioned concepts are verified, demonstrated and evaluated in various engineering and science applications [18],[19].

III. DQ METRICS COMPOSITION

Data may have various quality aspects, which can be measured. These aspects are also known as data quality dimensions, or metrics. Traditional dimensions are as follows, some of them are described in [20],[21]:

- **Completeness:** Data are complete if they have no missing values. It describes the amount, at which every expected characteristic or trait is described and provided.
- **Timeliness**: Timeliness describes the attribute that data are available at the exact instance of its request. If a user requests for data and is required to wait a certain amount of time, it is known as a data lag. This delay affects the timeliness and is not desirable.
- **Validity**: It determines the degree, at which the data conforms to a desired standard or rules.
- **Consistency**: Data are consistent if they are free from any contradiction. If the data conforms to a standard or a rule, it should continue to do so if reproduced in a different setting.
- **Integrity**: Integrity measures how valid, complete and consistent the data are. Data's integrity is determined by a measure of the whole set of other data quality aspects / dimensions.

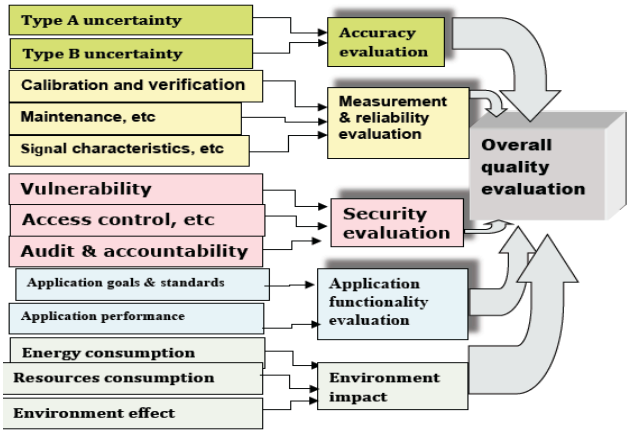

Figure 1. Integral quality evaluation composition

- **Accuracy**: Accuracy relates to the correctness of data and measurement uncertainty. Data with low uncertainty are correct.
- **Relevance**: It is a measure of the usefulness of the data to a particular application.
- **Reliability**: The quality of data becomes irrelevant if the data are not obtained from a reliable source. Reliability is a measure of the extent, to which one is willing to trust the data.
- **Accessibility**: It measures the timeliness of data.
- **Value added**: It is measured as the rate of usefulness of the data.

The methodologies of evaluating the DQ aspects listed above have been developed over the decades. They well represent the quality of the data at the point of their origin at the data source. However, nowadays most of the data are used far away from the point of their origin. In fact, the structured data are typically collected by distributed sensor networks and systems, then transmitted over the computer and communication networks, processed and stored by information systems, and, then, used. All those communication, processing and storage tasks affect the quality of data at the point of use, changing their DQ in comparison to one at the point of origin. The DQ evaluation should integrate accuracy and reliability of the data source with the security of the computer and communication systems. The high quality of the data at the point of their origin does not guarantee even an acceptable DQ at the point of use if the communication network security is low and the malicious alternation or loss of data has a high probability.

We describe the DQ evaluation structure as a multilevel hierarchical system. In this approach, we combine diverse evaluation systems, even if they vary in their design and implementation. The hierarchical system should be able to produce a partial evaluation of different aspects that will be helpful in flagging the areas that need urgent improvement. In our initial design we will classify metrics into five groups (see Figure 1):

 ISBN: 978-1-61208-376-6

TABLE I. SAMPLES OF GENERIC METRICS

Generic Attribute Name	DQ indicator/group (Figure1)	Description
Time-since-Manufacturing	Maintenance/reliability	The measure of the age of the device
Time-since-Service	Maintenance/reliability	The measure of the days since last service was performed in accord with the servicing schedule
Time-since-Calibration	Calibration/reliability	The measure of the days since last calibration was performed in accord with the calibration schedule
Temperature Range	Application/performance	The measure of temperature range within which the device will provide optimum performance
Physical Tampering Incidences	Physical security/security	The number of reported incidents that allowed unauthorized physical contact with the device
System Breaches	Access control/security	The measure of the number of unauthorized accesses into the system, denial of service attacks, improper usage, suspicious investigations, incidences of malicious code.
System Security	Security/security	Measures presence of intrusion detection systems, firewalls, anti-viruses.
Data Integrity	Vulnerabilities/securities	Number of operating system vulnerabilities that were detected.
Environmental Influences	Environment/environment	Number of incidences reported that would subject the device to mechanical, acoustical and triboelectric effects.
Atmospheric Influences	Environment/environment	Number of incidences reported that would subject the device to magnetic, capacitive and radio frequencies.
Response Time	Signals/reliability	Time between the change of the state and time taken to record the change

TABLE II. SAMPLES OF SPECIFIC DQ METRICS (EXAMPLES OF ELECTRIC POWER AND WATER METERS)

Device Name	Application specific Quality indicator	Description
Electric / Power Meters	Foucalt Disk	Check to verify the material of the foucalt disk.
	Friction Compensation	Difference in the measure of initial friction at the time of application of the compensation and the current friction in the device.
	Exposure to Vibrations	Measure of the number of incidences reported which would have caused the device to be subjected to external vibrations
Water Meters	Mounting Position	The measure of the number of days since regulatory check was performed to observe the mounting position of the device.
	Environmental Factors	Number of incidences reported which may have affected the mounting position of the device.
	Particle Collection	Measure of the amount of particle deposition.

(1) Accuracy evaluation;
(2) measurement and reliability evaluation;
(3) security evaluation;
(4) application functionality evaluation;
(5) environmental impact.

While the first three groups include rather generic metrics, groups #4 and #5 are devoted to metrics, which are specific to a particular application. Our metrics evaluation is based on existing approaches and standards, such as [22] for measurement accuracy and [23] for system security. Table I gives a sample of generic metrics representing all first three groups, while Table II lists the metrics, which are considered specific to a particular sensor and an application.

IV. DQ METRICS CALCULUS

In DQ calculus implementation we plan to investigate a wide number of options of calculating integral indicators from separate metrics ranging from simple weighted sums to sophisticated logical functions and systems. Those metrics and their calculation procedures will compose the DQ calculus. To simplify the calculus, we organize it as a hierarchical system calculating first the group indicators and then combining them into the system total. We follow the user-centric approach by offering an application user a choice of various options and their adjustment. We plan to introduce a function choice automatic adjustment,

 ISBN: 978-1-61208-376-6

verification and optimization.

To realize a wide variety of logical functions, the expert system technology is employed as the main implementation technique. The automated tool set includes the hierarchical rule-based systems deriving values for separate metrics, then combining them into groups and finally producing an overall evaluation. This way, the tool operation follows up the metrics structure and composition (see figure 2). This system needs to be complemented by the tools and databases assisting automation of all stages in the data collection, communication, processing and storage for all information available for data quality evaluation. The developed tools facilitate automated collection, communication and processing of the relevant data. Based on the data collected, they not only evaluate the overall data quality but also determine whether or not the data collection practice in place is acceptable and cite areas that are in need of improvement.

In our automated procedures, the DQ score is computed by applying either linear, exponential or stepwise linear reduction series to the maximum score of an attribute. In case an attribute defines a range for ideal working, the linear series is substituted by a trapezoidal drop linear series and exponential is replaced by a bell drop series.

When considering both accuracy and security DQ metrics, assessing whether fusion enhances DQ is not obvious as one has to tradeoff between accuracy, security and other goals. While adding up a more secure data transmission channel improves both security and accuracy indicators, using a more accurate data stream will definitely improve data accuracy but could be detrimental to certain security indicators (see [24] for further discussion). If resources are limited, as in the case of sensor networks, one might consider trying to improve accuracy of the most secure data source versus more or less even distribution of security resources in order to achieve the same security levels on all data channels. The concrete recommendations will depend on the application.

V. GENERIC TOOL DESIGN

The proposed design of the tool divides the procedure for automated data collection in three main stages. First stage involves mainly a device configuration. Since the tool is generic, it provides certain flexibility in configuring a large variety of diverse devices. These devices could be electric meters, power meters, water meters and marine sensors. The second stage computes data quality indicators of the configured device. The final stage performs the detailed analysis of the computed data quality indicators. It highlights low data quality and help flag erroneous data. Also, it provides recommendations on improving low data quality and help ensure that the data being utilized are fit for the purpose it is intended to be used. Figure 2 presents the architecture of the tool. Currently, the first and the second stages are implemented.

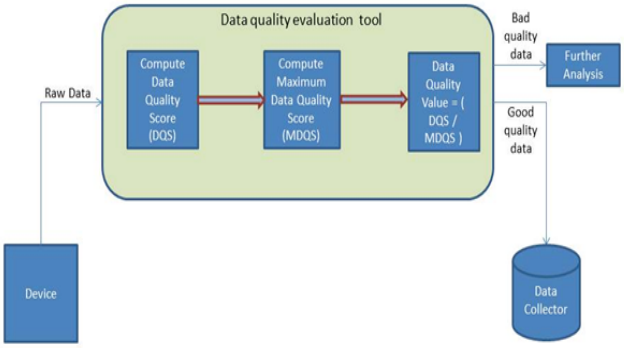

Figure 2. Data quality evaluation procedure

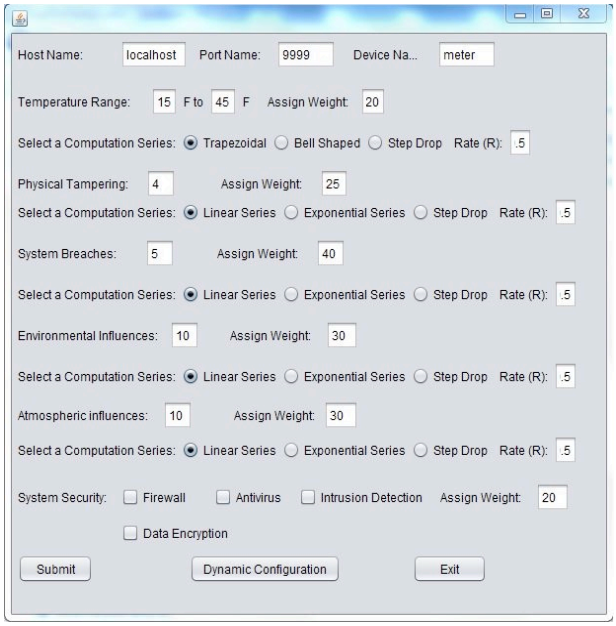

Figure 3. Generic attribute configuration

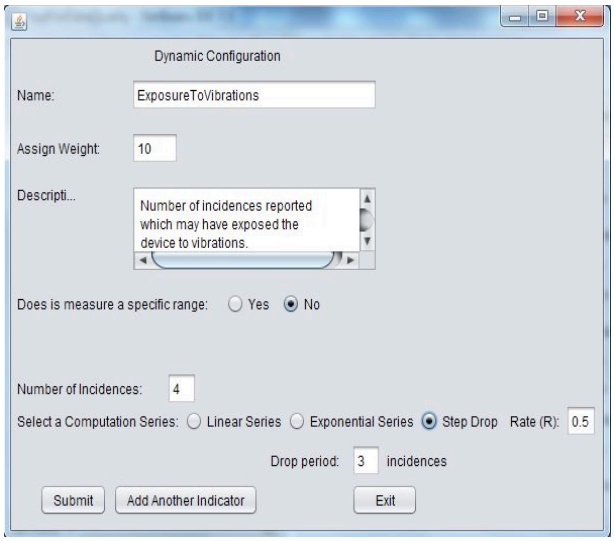

Figure 4. Application specific attribute configuration

The generic tool allows for a configuration of a large variety of devices. Each automated data collection device has DQ factors, which are common to other similar devices. These factors are referred by the tool as generic attributes. Other attributes, which are unique to a particular device are called dynamic attributes. These attributes are assigned the maximum score based on the significance of the contribution they would add to the data quality. The greater the significance, the greater is the score.

The configuration step mainly involves recognizing the generic and application-specific attributes, as well as assigning the max possible score to each of them. Generic attributes are common to most devices, for example, timeliness and quality of common device servicing such as calibration. Application-specific attributes are unique to a device, for example, exposure to vibration, shock and radiation. This is important for a particular application because certain devices, like electric meters, produce misleading results when exposed to the external adversary affects. If, for some reason, a generic attribute does not apply to a particular device, the max score of zero would be applied in order to eliminate the attribute from the analysis. Table I describes the generic attributes being considered by the tool. Figure 3 illustrates configuring some of the generic attributes for an electric meter. Table II describes some application specific attributes, which are device and application specific. Figure 4 illustrates configuring an application-specific attribute for an electric meter, provided as an example.

The second stage involves data quality computation. The configured generic and application specific attributes help compute the individual quality scores. Each attribute is considered a quality indicator, whose significance will be dependent on its max score. These quality indicators produce a quality score using a chosen logic procedure. For example, we can consider a generic attribute called time-since-calibration. Some devices need to get calibrated every year. If a device has not been calibrated for an entire year or a couple of years, the quality factor for that indicator will go down. If the device has never been calibrated since its installation it can affect the quality score even more. The tool allows a user to define the procedure for calculating the application-specific quality indicators.

VI. CONCLUSIONS

The paper introduces a novel approach to data management in data collection and processing systems, which might incorporate SCADA, sensor networks and other systems with nanoscale devices. We associate each data entity with the corresponding DQ indicator. This indicator integrate various data characteristics ranging from accuracy to security, privacy and safety, etc. It considers various samples of DQ metrics representing communication and computing security as well as data accuracy and other characteristics. Incorporating security and privacy measures into the DQ calculus is especially important in the current development as it allows shifting the DQ assessment from the point of data origin to the point of data use.

A unified framework for assessing DQ is critical for enhancing data usage in a wide spectrum of applications because this creates new opportunities for optimizing data structures, data processing and fusion based on the new DQ information use. By providing to an end user or an application the DQ indicators which characterize system and network security, data trustworthiness and confidence, etc. Correspondingly, an end user is in a much better position to decide whether and how to use data in various applications. A user will get an opportunity to understand and compare various data files, streams and sources based on the associated DQ with integral quality characteristics reflecting various aspects of system functionality and to redesign data flows schemes. This development will transform one-dimensional data processing into multi-dimensional data optimization procedures for application-specific data applications. We describe and demonstrate an application of the DQ metrics definition and calculation tools, which enable integration of various metrics to calculate an integral indicator.

REFERENCES

[1] P. W. Coteus, J. U. Knickerbocker, C. H. Lam and Y. A. Vlasov, “Technologies for exascale systems,” IBM Journal of Research and Developments, vol. 55, issue 5, pp. 14.1-14.12, 2011.

[2] *Handbook on Nano and Molecular Electronics*, Ed. S. E. Lyshevski, CRC Press, Boca Raton, FL, 2007.

[3] B. G. Lee et. al., “Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits,” Journal of Lightwave Technology, vol. 32, issue 4, pp. 743-751, 2014.

[4] S. E. Lyshevski, Molecular Electronics, Circuits and Processing Platforms, CRC Press, Boca Raton, FL, 2007.

[5] Micro-Electromechanical Systems (MEMS), International Technology Roadmap for Semiconductors, 2011 and 2013 Editions, available at www.itrs.net , accessed on August 1, 2014.

[6] A. Yariv, *Quantum Electronics*, John Wiley and Sons, New York, 1988.

[7] Emerging Research Devices, International Technology Roadmap for Semiconductors, 2011 and 2013 Editions, available at www.itrs.net , accessed on August 1, 2014.

[8] J. Warnock, “Circuit and PD challenges at the 14nm technology node,” Proc. 2013 ACM Int. Symposium on Physical Design, pp. 66-67, 2013.

[9] M. Klein and W. Lehner, "Representing Data Quality in Sensor Data Streaming Environments," J. Data and Information Quality, vol. 1, pp. 1-28, 2009.

[10] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering Data Streams: Theory and Practice," IEEE Trans. on Knowl. and Data Eng., vol. 15, pp. 515-528, 2003.

 ISBN: 978-1-61208-376-6

[11] H. S. Lim, K. M. Ghinita, and E. Bertino "A Game-Theoretic Approach for High-Assurance of Data Trustworthiness in Sensor Networks," presented at the IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA, 2012.

[12] C. Dai, H.S. Lim, and E. Bertino "Provenance-based Trustworthiness Assessment in Sensor Networks," , 7th Workshop on Data Management for Sensor Networks (DMSN), in conjunction with VLDB, DMSN 2010, Singapore, 2010.

[13] S. Chobsri, W. Sumalai, and W. Usaha, "Quality assurance for data acquisition in error prone WSNs," in Ubiquitous and Future Networks, 2009. ICUFN 2009. First International Conference on, 2009, pp. 28-33.

[14] W. Dong, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, and C. C. Aggarwal, "Optimizing quality-of-information in cost-sensitive sensor data fusion," in 2011 International Conference on Distributed Computing in Sensor Systems (DCOSS 2011), 27-29 June 2011, Piscataway, NJ, USA, 2011, 8 pp.

[15] W. Yinfeng, W. Cho-Li, C. Jian-Nong, and A. Chan, "Optimizing Data Acquisition by Sensor-channel Co-allocation in Wireless Sensor Networks," in 2010 International Conference on High Performance Computing (HiPC 2010), 19-22 Dec. 2010, Piscataway, NJ, USA, 2010, p. 10 pp.

[16] L. Reznik, "Integral Instrumentation Data Quaity Envaluation: the Way to Enhance Safety, Security, and Environment Impact," 2012 IEEE International Instrumentation and Measurement Technology Conference, Graz, Austria, May 13-16, 2012, 2012.

[17] S. E. Lyshevski and L. Reznik, "Processing of extremely-large-data and high-performance computing," in International Conference on High Performance Computing, Kyiv, Ukraine, 2012, pp. 41-44.

[18] G. P. Timms, P. A. J. de Souza, L. Reznik, and D. V. Smith, "Automated Data Quality Assessment of Marine Sensors," *Sensors,* vol. 11, pp. 9589-9602, 2011.

[19] G. P. Timms, P. A. de Souza, and L. Reznik, "Automated assessment of data quality in marine sensor networks," in OCEANS 2010 IEEE - Sydney, 2010, pp. 1-5.

[20] F. G. Alizamini, M. M. Pedram, M. Alishahi, and K. Badie, "Data quality improvement using fuzzy association rules," in Electronics and Information Engineering (ICEIE), 2010 International Conference On, 2010, pp. V1-468-V1-472.

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

[22] ANSI/NCSL, "US Guide to the Expression of Uncertainty in Measurement," ed, Z540-2-1997.

[23] National Institute of Standards and Technology, "Performance Measurement Guide for Information Security," ed. Geithersburg, MD, July 2008.

[24] L. Reznik and E. Bertino, "Poster: Data quality evaluation: integrating security and accuracy," Proceedings of the 2013 ACM SIGSAC conference on Computer communications security, Berlin, Germany, 2013.

 ISBN: 978-1-61208-376-6