5D-IoT, a semantic web based framework for assessing IoT data quality

Nathalie Jane Hernandez, Shubham Mante, Aftab M Hussain, Sachin Chaudhari, Deepak Gangadharan, Thierry Monteil

▶ **To cite this version:**

Nathalie Jane Hernandez, Shubham Mante, Aftab M Hussain, Sachin Chaudhari, Deepak Gangadharan, et al.. 5D-IoT, a semantic web based framework for assessing IoT data quality. 37th ACM/SIGAPP Symposium on Applied Computing (SAC 2022), Apr 2022, Virtual Event, France. pp.1921-1924, 10.1145/3477314.3507234 . hal-04279670

HAL Id: hal-04279670
https://hal.science/hal-04279670

Submitted on 10 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

5D-IoT, a Semantic Web Based Framework for Assessing IoT Data Quality

Shubham Mante
shubham.mante@research.iiit.ac.in
International Institute of Information Technology
Hyderabad, Telengana, India

Nathalie Hernandez
hernande@irit.fr
IRIT, UT2J, University of Toulouse
Toulouse, France

Aftab M Hussain
aftab.hussain@iiit.ac.in
International Institute of Information Technology
Hyderabad, Telengana, India

Sachin Chaudhari
sachin.c@iiit.ac.in
International Institute of Information Technology
Hyderabad, Telengana, India

Deepak Gangadharan
deepak.g@iiit.ac.in
International Institute of Information Technology
Hyderabad, Telengana, India

Thierry Monteil
thierry.monteil@irit.fr
IRIT, INSA, University of Toulouse
Toulouse, France

ABSTRACT

Due to the increasing number of Internet of Things (IoT) devices, a large amount of data is being generated. However, factors such as hardware malfunctions, network failures, or cyber-attacks affect data quality and result in inaccurate data generation. Therefore, to facilitate the data usage, we propose a novel 5D-IoT framework for heterogeneous IoT systems that provides uniform data quality assessment with meaningful data descriptions. Based on the quality assessment result, a data consumer can directly access data from any IoT source, which ultimately speeds up the analysis process and helps gain important insights in less time. The framework relies on semantic descriptions of sensor observations and SHACL shapes assessing the quality of such data. Evaluations carried out on real-time data show the added value of such a framework.

KEYWORDS

IoT data quality assessment, Semantic web of things, SHACL, SPARQL

ACM Reference Format:
Shubham Mante, Nathalie Hernandez, Aftab M Hussain, Sachin Chaudhari, Deepak Gangadharan, and Thierry Monteil. 2022. 5D-IoT, a Semantic Web Based Framework for Assessing IoT Data Quality. In *The 37th ACM/SIGAPP Symposium on Applied Computing(SAC '22), April 25–29, 2022, Virtual Event,*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3477314.3507234

1 INTRODUCTION

Data is now openly available, thanks to the emerging IoT projects that provide data with the help of Web Application Programming Interfaces (APIs). However, often open data from such data streams becomes difficult to process and analyze due to the obscure key-value pairs and incomplete or unclear data semantics. The Semantic Web (SW) technologies and principles solves these issues to improve interoperability by providing expressive vocabularies to describe data and manipulate resources. IoT systems faces other problems such as additional data transmission latency, duplicate data generation, erroneous and null data transmission [11]. Therefore, to ensure the quality of the IoT data, these factors needs to be validated. To deal with data validation the World Wide Web Consortium (W3C) recommends the Shapes Constrained Language (SHACL) [9]. In [5], authors proposed an annotation engine called LSane, which semantically enriches the data and validates it using SHACL constraints before reaching the customers. Authors mentioned about the *shapes* that consider factors such as data attributes, cardinality of relations or datatype restrictions to ensure data quality. However, authors did not discuss about other factors such as completeness, accuracy and timeliness which also affect data quality and hence needs to be validated. There arises a need of a framework that provides the data with uniform data semantics and quality assessment information. Therefore, in this paper, we propose a five-layered 5-D IoT framework, namely data enrichment, data duplicacy, data delay, data validation, and data storage. The remainder of the paper is divided into three sections. The core contribution is detailed in section 2, and section 3 describes a prototype implementation of the framework and its evaluation. This paper is concluded in section 4.

2 THE PROPOSED METHODOLOGY

Our aim is to provide explicit semantics with quality assessment information to facilitate the use of data. In this section, we describe the proposed 5D-IoT framework that enriches the meaning of the data and evaluates it for quality assessment. An overview of the proposed framework is illustrated in Fig. 1. In the first layer, the raw data from heterogeneous IoT sources is uniformly enriched and converted into graph data based on the Resource Description Framework (RDF) [3][1]. Such enriched data makes it easy for data consumers to understand the meaning of data leading to faster application development. The next three layers represent the novel approach of the data quality assessment using SHACL. *shapes*, a collection of constraints applied to specific RDF resources are used to assess duplicate data transmission, data transmission delays, and inaccurate data transmission, respectively. The assessment results are added to the data graph using the SHACL inference rule and

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SAC '22, April 25–29, 2022, Virtual Event,
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8713-2/22/04.
https://doi.org/10.1145/3477314.3507234

[1] https://www.w3.org/RDF/

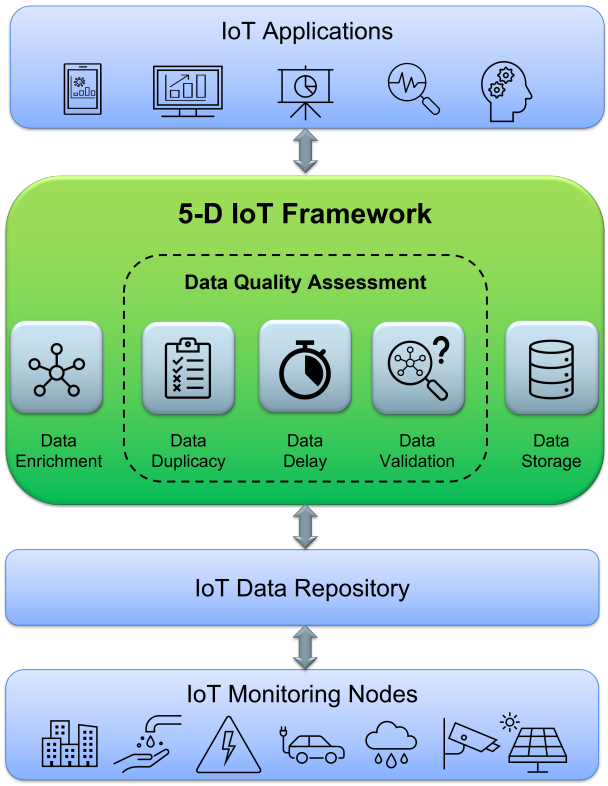

Figure 1: Illustration of the five-layered 5-D IoT framework

our proposed vocabulary. Finally, the fifth layer consists of the RDF Triple Store, which stores and provides the processed data to users.

2.1 Data Enrichment

In the first layer of the framework, i.e., in the data enrichment layer, the SOSA-SSN ontology [7] is used as the primary ontology to describe the data. User modifiable templates are considered to map a new or latest observation to its associated semantics to convert it into an RDF graph. The observation is described with the observed property (sosa: observedProperty), the entity whose property is being observed (sosa: hasFeatureOfInterest), the observed value (sosa: hasResult), its datatype, unit (qudt-unit-1-1: QualityValue), time of observation (sosa: resultTime) and the sensor that made the observation (sosa: madeBySensor). In this way, the framework enriches the data and forwards it to the next layer for quality assessment. The Fig. 3, shows an example of data description of a temperature observation using an **ex** prefix.

ex:	http://example.org/data/
idqa:	http://example.org//2021/01/idqa#
sosa:	http://www.w3.org/ns/sosa/
ssn:	http://purl.oclc.org/NET/ssnx/ssn#
xsd:	http://www.w3.org/2001/XMLSchema#
qudt-1-1:	http://qudt.org/1.1/schema/qudt#
qudt-unit-1.1:	http://qudt.org/1.1/vocab/unit#
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#

Figure 2: Ontology prefixes for Fig. 3 and 4

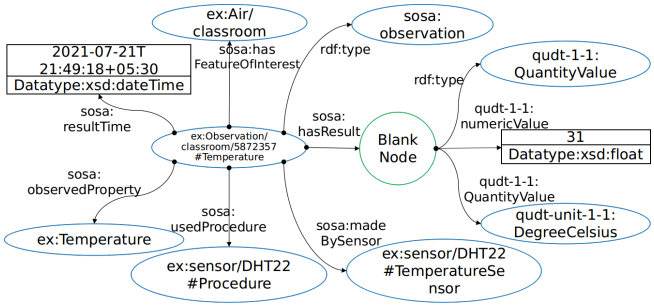

Figure 3: Graph describing a temperature sensor observation

2.2 Data Quality Assessment

IoT systems receive data from heterogeneous devices deployed in multiple locations that monitor numerous properties, and the values of these properties change over time; hence, it becomes challenging to evaluate the quality of IoT data. Therefore, we propose a dynamic mechanism by using SHACL-SPARQL [13] *shapes* to assess the data by collectively considering the feature of interest (FOI), the property that is being observed, the observed value, and time of the observation. Each shape is built using a template that is based on the quality factors described in the knowledge base with the vocabulary presented above. A *shape* can be activated or deactivated based on the framework's configuration. It provides the flexibility to the framework administrator to apply various constraints without modifying the whole framework.

2.2.1 IDQA Vocabulary: The proposed vocabulary describes two aspects: the factors that should be considered while assessing the quality and the quality assessment result.

For the first aspect, we define three classes and six properties. The class **idqa: QualityFactor** represents an observation quality factor that can depend on a given observable property at a given interval of time for a given FOI. To represent such information the data property **idqa: forInterval** and the two object properties, **idqa: forFeatureOfInterest** and **idqa: forObservableProperty** are proposed. Moreover, two types of quality factors are defined as subClasses of **idqa: QualityFactor**; **idqa: ExpectedDelay** and **idqa: RangeValue**. For a **idqa: ExpectedDelay** factor, the expected delay between two successive observations for a given feature of interest and a given property is defined through the data property **idqa: delayValue**. For a **idqa: rangeValue** factor, two data properties are proposed; a minimum expected observation value for the results of such observations that is expressed through the data property **idqa: minValue**, and a maximum expected observation value expressed through the data property **idqa: maxValue**. Fig. 4 illustrates the use of this vocabulary on two quality factors.

For the second aspect, which represents the quality assessment of the result, we define four classes and four properties. We reuse the object property system:qualityOfObservation from the sosa-ssn ontology and define a subproperty **idqa:qualityOfObservation** linking a sosa:Observation to **idqa:QualityAssessment**. The latter has three sub-classes: **idqa: DuplicacyAssessment**, **idqa: DelayAssessment** and **idqa: RangeValueAssessment**. The class **idqa: DuplicacyAssessment**, describes the potential duplicate packets received with the **idqa: numOfDuplicates** data property. By using **idqa: samplingDelay** and **idqa: transmissionDelay**

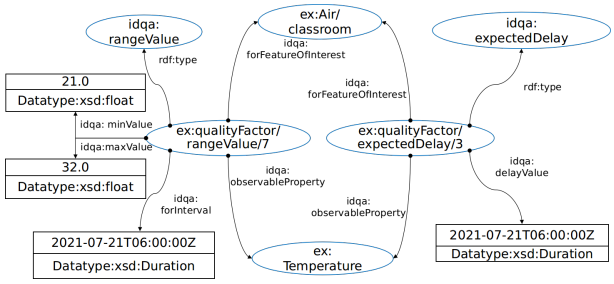

Figure 4: Graph describing two quality factors

data properties, the class **idqa: DelayAssessment** describes both the sampling time delay caused by the IoT node while collecting the data and the transmission delay caused by the transmission of data from an IoT node to the data storage platform. Finally, with the help of **idqa: isOutofRange** data property, the class **idqa: RangeValueAssessment** describes whether the observation value is within the **idqa: minValue** and **idqa: maxValue**.

These quality descriptions will be then added to the observation graph to help data consumers and data providers understand the assessment results and, based on that, improve the data quality.

2.2.2 Assessment Techniques:

2.2.2.1 Data Duplicacy: As mentioned in [10], one of the most common problems in IoT systems is the transmission of duplicate data, i.e., multiple transmissions of the same measurement value. Duplicate data can be identified by comparing the timestamps of two successive non-duplicate observations. In the proposed framework, a shape using the SHACL-SPARQL constraint is applied to every observation. The shape validates if the timestamp of the newly received observation (t_{o_new}) is less than or equal to that of the previous non-duplicate observation (t_{o_last}) from the same FOI and observable property. Based on the validation result, an inferred triple will be added to the observation graph using the CONSTRUCT-based SHACL advanced rule. So, if it is a non-duplicate observation, then 0 is added as an object value to the data property, **idqa: numOfDuplicates** otherwise, n, the number of times observation is received will be inferred and added.

2.2.2.2 Data Delay: Sometimes, network failures and hardware malfunctions can cause an IoT node to send delayed data, which affects the performance of applications that receive such delayed data. As a first step, we divide the delay problem into two parts, transmission delay, and sampling delay. The transmission delay is caused during the transmission of data from node to the IoT platform, and sampling delay is caused during the collection of data by an IoT device (sosa: Sampler). The transmission delay is calculated by taking a difference of observation time (t_{o_new}) and the time when the observation is recorded in the data repository platform (t_r). To calculate the sampling delay, first, the timestamp of the last non-duplicate observation (t_{o_last}) is queried from the RDF triple-store. Using the FILTER SPARQL query of SHACL-SPARQL constraint, that timestamp is compared with the latest observation's timestamp (t_{o_new}). If the difference between the timestamps exceeds the **idqa: delayValue** (T) stored for the instance of the class **idqa: ExpectedDelay** linked to the corresponding FOI and observable property, it means the received observation is delayed. Later, triples that contain the delay values are inferred and added to the observation graph using the data properties, **idqa: transmissionDelay** and **idqa: samplingDelay**.

2.2.2.3 Data Validation: Devices in IoT systems observe various properties whose values differ over various factors like time, location, the FOI or device capabilities. We propose a novel approach to use the SHACL shapes, which apply dynamic range value constraints for data validation. When a new, semantically enriched, non-duplicate, delayed or non-delayed observation is received at the data validation layer; first, its FOI and observable property is checked. Based on these values, the minimum (V_{min}) and maximum (V_{max}) values represented with the data properties **idqa: minValue idqa: maxValue** are retrieved from triple store and used while applying the range value constraint to validate the outliers. So, if the observation value ($V_{observation}$) is null or out of the allowed range, it is declared as inaccurate data, and an inferred triple is added to the observation graph using **idqa: isOutofRange** data property with value as TRUE.

2.3 Data Storage

The fifth and final layer consists of an RDF triplestore capable of storing the completely processed data. This stored data will benefit various applications such as quality assessed data archiving, descriptive data visualization, correct data analysis and many more.

3 EVALUATION

3.1 Implementation

The proposed framework and the vocabulary are available online[2]. The framework is implemented using Django [15], an open-source framework for backend web applications based on Python. The APIs provided by the Django REpresentational State Transfer (REST) framework enable the data transmission from repositories to the framework and framework to end-users. Apart from built-in libraries of Python, other libraries were used, such as RDFLib [3] to work with RDF in Python and pySHACL [16] to validate RDF data graphs using SHACL shapes. We have assumed that the data packet received by the platform would be a JavaScript Object Notation (JSON) object containing a set of key-value pairs with or without any semantics. The Apache Jena Fuseki [8] triple store is used to store the processed RDF data. In the first step, sensor graphs, features of interests, expected range value constraints, the expected delay between two successive observations, and templates for shapes are loaded and stored. Then, once the framework receives a new observation, its values are parsed and enriched by linking them with associated properties and converted into an RDF graph using the RDFLib library. Every new observation gets a unique resource identifier (URI) that helps while adding the inferred triples to the observation graph. The RDF graph is then validated against SHACL shapes using the pySHACL module. Based on the validation result, the data properties for sosa: Observation will be added to the observation graph to describe the quality assessment of an observation. Finally, the completely processed data is stored in Fuseki.

[2]https://github.com/shubham-mante/5D-IoT-framework
[3]https://rdflib.readthedocs.io/en/stable/

Table 1: IoT Data Quality Assessment Results

IoT Data Monitoring Applications	*IoT Application Characteristics*				
	Expected Observations	***Received Observations***	***Duplicate Observations***	***Delayed Observations (Transmission/Sampling)***	***Inaccurate Observations***
Energy Usage	96	76	29	47/23	0
Water Flow	480	365	0	365/365	365 (Pressure Voltage)
Weather Condition	1440	1440	556	36/36	365 (Temp,Humid,Wind-Speed)
Air Quality	5760	3473	1726	1673/1673	3378 (PM2.5, PM10)

3.2 Results

The results detailed in Table 1 are obtained by evaluating the real-time monitoring data received from the Smart city living lab, IIIT-H repository [4]. The repository collects data from heterogeneous IoT verticals such as air [14], water [6], energy [12], and weather monitoring. Each devices monitors multiple parameters, for example, air quality device monitor temperature, humidity, PM10, PM2.5, and CO2 values. We have processed 24-hours data from four different monitoring devices to evaluate the framework. The devices are deployed in different working conditions and transmit data at different sampling periods according to application requirements. During the duplicacy assessment, we observed that the duplicate data transmission for each device differs due to the difference in sampling period and working conditions. For example, sometimes, a device with low sampling period does not receive an acknowledgment from the IoT platform, and hence tries to re-transmit the same observation until it receives the acknowledgment. Due to such repetitive transmission, the device misses the transmission of new observations and as a result data is lost. Based on the delay assessment results, we found out that the observations received over 24 hours from the air quality monitoring node produces a significant amount of delay. Such delays result in not only data loss but also false results prediction if it is done based on the ideal sampling period set by the data provider. It is observed that the cause of inaccurate data transmission is different for every IoT node. Other hardware malfunction issues were found out that resulted in the transmission of data beyond the allowed range. The results detailed in Table 1 illustrate the number of received observations of the observed properties whose values are out of the allowed range.

4 CONCLUSION AND FUTURE SCOPE

This paper proposes a novel framework to assess the data quality using SHACL shapes. It explicitly provides the quality assessment information as a part of the description of the observations. To the best of our knowledge, the current data quality ontologies such as DQV [1], and DaQ [4], either focus on assessing the quality of a dataset or focus on a specific domain. These ontologies can be used to assess the quality of the dataset formed by the 5D-IoT framework, but as our primary focus is to assess each observation, we did not use any of them. Later in the future work, the aim is to evaluate the added value of data filtering according to assessed quality for consumers deployed on the network and automatic generation of SHACL *shapes* similar to ASTREA [2] to eliminate the need for expert knowledge for *shapes* creation and to use more constraints for assessing the IoT data quality.

[4]https://smartcityresearch.iiit.ac.in/smartcampus.html

ACKNOWLEDGMENT

This research was partly supported by the Ministry of Electronics and Information Technology (MEITY) under grant no. 3070665 (2020) as part of the Smart City Living Lab project.

REFERENCES

[1] Riccardo Albertoni and Antoine Isaac. 2021. Introducing the Data Quality Vocabulary (DQV). *Semantic Web* Preprint (2021), 1–17.
[2] Andrea Cimmino, Alba Fernández-Izquierdo, and Raúl García-Castro. 2020. Astrea: automatic generation of SHACL shapes from ontologies. In *European Semantic Web Conference*. Springer, 497–513.
[3] Richard Cyganiak, David Wood, and Markus Lanthaler. 2015. Rdf 1.1 concepts and abstract syntax, 2014. *URL https://www. w3. org/TR/rdf11-concepts* (2015).
[4] Jeremy Debattista, Christoph Lange, and Sören Auer. 2014. daQ, an ontology for dataset quality information. In *LDOW*.
[5] Matthias T Frank, Sebastian Bader, Viliam Simko, and Stefan Zander. 2018. Lsane: Collaborative validation and enrichment of heterogeneous observation streams. *Procedia Computer Science* 137 (2018), 235–241.
[6] Sai Usha Nagasri Goparaju, SVSLN Surya Suhas Vaddhiparthy, C Pradeep, Anuradha Vattem, and Deepak Gangadharan. 2021. Design of an IoT System for Machine Learning Calibrated TDS Measurement in Smart Campus. In *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*. 877–882. https://doi.org/10.1109/WF-IoT51360.2021.9595057
[7] Krzysztof Janowicz, Armin Haller, Simon JD Cox, Danh Le Phuoc, and Maxime Lefrançois. 2019. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics* 56 (2019), 1–10.
[8] Apache Jena. 2014. Apache jena fuseki. *The Apache Software Foundation* 18 (2014).
[9] Holger Knublauch and Dimitris Kontokostas. 2017. Shapes Constraint Language (SHACL), W3C Recommendation 20 July 2017. *URL: https://www. w3. org/TR/shacl* (2017).
[10] Hua Li, Huan Wang, Wenqing Yin, Yongwei Li, Yan Qian, and Fei Hu. 2015. Development of a remote monitoring system for henhouse environment based on IoT technology. *Future Internet* 7, 3 (2015), 329–341.
[11] Taha Mansouri, Mohammad Reza Sadeghi Moghadam, Fatemeh Monshizadeh, and Ahad Zareravasan. 2021. IoT Data Quality Issues and Potential Solutions: A Literature Review. *arXiv preprint arXiv:2103.13303* (2021).
[12] Shubham Mante, Ruthwik Muppala, D. Niteesh, and Aftab M. Hussain. 2021. Energy Monitoring Using LoRaWAN-based Smart Meters and oneM2M Platform. In *2021 IEEE Sensors*. 1–4. https://doi.org/10.1109/SENSORS47087.2021.9639822
[13] E Prud'hommeaux. 2015. SHACL-SPARQL (W3C Editor's Draft). W3C.
[14] C. Rajashekar Reddy, T. Mukku, Ayush Dwivedi, Ashrit Rout, Sachin Chaudhari, Kavita Vemuri, Krishnan S. Rajan, and Aftab M. Hussain. 2020. Improving Spatio-Temporal Understanding of Particulate Matter using Low-Cost IoT Sensors. In *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. 1–7. https://doi.org/10.1109/PIMRC48278.2020.9217109
[15] Daniel Rubio. 2017. REST services with Django. In *Beginning Django*. Springer, 549–566.
[16] Ashley Sommer, Piyush Gupta, Nolan Nichols, Nicholas Bollweg, Alex Nelson, James Howison, Nicholas Car, Jamie Feiss, Jonathan Yu, Mohameth François SY, and et al. 2021. RDFLib/pySHACL: RDFLib 6.0.0 Support. (Jul 2021). https://doi.org/10.5281/zenodo.5115668