Data Quality Computation For Obsolescence Detection Within Connected Environments

Jean Raphael Richa

Data Quality Computation For Obsolescence Detection Within Connected Environments

Jean Raphael Richa
University Pau & Pays Adour
LIUPPA
Anglet, France
jean-raphael.richa@etud.univ-pau.fr

***Abstract*—The unceasing growth in digital technologies has promoted the means of sensing, visualizing, and autonomously analyzing the environment encompassing us. These connected environments provide data interoperability producing and collecting a staggered quantity of information deemed forefront in better understanding the challenges our societies face. This has helped in enhancing the performance of new or existing entities with product life-cycle and smart cities being just two examples. However, making use of data is one perspective, and being able to detect which data is actually useful is a different perspective. The latter encounters a critical research gap as there is no clear procedure for identifying obsolete data. Therefore, this paper aims to clearly identify data quality metrics purposed for data obsolescence detection within a connected environment.**

***Index Terms*—Connected environments, Data obsolescence, Data Quality Metrics.**

I. Introduction

Living in the fast tech era, the focus is mainly on data security, interoperability, and speed with less emphasis on the data integrity; namely, the data that is no longer useful. In a world itself being a fast-paced environment changing faster than our ability to comprehend, comes premature obsolescence [2] leading to unsustainability in information and communication technologies. Data obsolescence can have different meanings based on the context it's involved in. Focusing on the information technology sector, it is defined as antiquated data gone/going out of use, data no longer practical, or data not needed to begin with [6].

Obsolescence is a major cause of high cost and integrity loss; nonetheless, it is often not taken into consideration as a serious issue nor included within the planning and implementation stages [6]. With the gradual loss of data confidentiality runs a risk of erroneous behavior and false data representation [9]. For instance, this causes the United States defense and aerospace sector to face an approximate loss of 750$ million annually according to the US Navy [19]. Accordingly, it is crucial to adopt a goal-oriented approach to minimize the impact of obsolescence on organizations especially for communication and sensing industries. Smart cities are increasingly integrated with various devices, systems, and sensors within a networked ecosystem; connected environments, allowing them to communicate, share, and collaborate on data seamlessly. However, with vigorous data generation comes again the need for new data integrity measures. Consequently, being the first representatives of this initiation, we will unveil which data quality parameters to use alongside computation techniques providing the basic infrastructure needed for obsolescence detection in connected environments.

Research is aiming to adequately manage, reduce and solve the negative impacts of data obsolescence; however, there is currently still no clear definition as to what data obsolescence is and how it corresponds to the new ongoing technical revolution. Its existence cannot be narrowed down only to the aspect of data no longer used; there are many more reasons behind the cause of it. It is to be a stand-alone parameter with its own definition and dimensions; the topic this paper aims to uncover in the field of connected environments. In this paper, we aim to provide data obsolescence with a clear set of metrics and definitions purposed in its detection within connected environments.

II. State of the Art

Data obsolescence can exist almost in any topic and can have various definitions based on the context it is used in. In [12], the authors focus on data obsolescence in the domain of products such as when a product becomes non-functional prematurely with a shorter life span. The authors argue that targeting obsolescence policies and interventions can contribute to more sustainable technology. In [5], the authors also focus on both the product, its users, and how they are affected by what they called "Technological Obsolescence". The latter afflicts almost every developed product claimed, some of which are slowly affected (e.g. roads, bridges) while others rapidly (e.g. audio industry). The authors in [5] further extend this to a functional definition of obsolescence describing it as a road becoming obsolete just because the destination is no longer in use. In other words, due to technological advancements, the market trends shift into something new making the old trend antiquated. The authors of [20] also discuss the obsolescence of products/services, defining it as the aging of a product in means of the production process and materials. They consider a product's obsolescence can be one of four types: material, functional, psychological, or economical. While the authors managed to describe four obsolescence types, these types are limited to the product as a physical entity. In this context, the product life cycle management of data centers was the most important factor here and not treating the data centers

themselves; one of the gaps this research paper aims to show. Examining life cycles as well, extending that of defense and aerospace sectors through managing obsolescence is covered by the authors of [18]. Knowingly, aerospace uses Concept, Assessment, Demonstration (CAD) phases which can last as long as 10 years as the authors explained. This is a major issue as obsolescence may arise even before the actual development phase starts. This suggests the seriousness of data obsolescence and the need of its management in early stages to reduce the associated risks.

Contemporary Internet of Things (IoT) systems are producing fast growing volumes of data. Accordingly, while the new data is coming into storage, previously saved data becomes excessive with obsolete data. As it's unprofitable to store all original data, in [1], the authors investigate obsolescence in data storage while making use of aggregation methods to minimize the amount of it stored. Their methodology is limited to two factors, old data versus current data, whereby the latter is represented as the system's latest state. Data is then labeled as obsolete based on its probability of necessity; any data no longer requested becomes obsolete. While expressive in their application, the authors do not comprehensively consider obsolescence as their focus is more on when the sensors have collected the data. However, sensors, whether real or virtual, collect a large set of data on a real time basis, as the authors of [17] foreground. With the digital twin becoming a growing trend, the authors develop a connected environment of a mimicked real world composed of virtual sensors and weather phenomenon to measure. Although the authors do not provide data analysis, they emphasize on the importance of the data to be collected by the sensors. Accordingly, understanding such data while detecting which is useful and which is obsolete, will help handle spatial–temporal redundancies and enhance systems' processing time. Knowingly, programmed environments intend to grow indefinitely while the users wait on standard backup discs to archive data, the authors of [11] highlighted. The authors stress on the importance of removing obsolete data in order to prevent data explosion, and how it "vanishing" is a peace of mind in business. While this reflects the research attention data obsolescence is in need of, it also highlights the lack of perception this topic has. Namely, the aim is not to vanish obsolete data on the spot, but rather understand first the data itself then act accordingly. [8] expands the importance of obsolescence highlighting that even new data can contradict existing ones. The authors focus on the contradiction detection accuracy variable "-Contradiction" to estimate any obsolescence. However, the scope was limited to two consistent databases compromising the same subject of information with criteria considered to be "perfect and fixed during processing". Also covering databases are authors of [9] who describe how the confidence in data decays with time as we are held behind the world's fast-paced evolution. Understanding the importance of data obsolescence, the authors aim to reduce the problems faced through monitoring and controlling obsolescence. Although they categorize obsolescence metrics minimalistically, the authors highlight the amount of research needed in this area and the need for more data credibility.

In [4], the authors explain how information systems' success is closely related to data freshness. Even though the authors do not discuss obsolescence, they describe quality metrics (such as currency/materialization, staleness, timeliness) representing freshness level of data. Knowingly, freshness is not only about the latest data received, it's also how important that data is. For instance, old data may have a quality dimension deeming it necessary to keep while a newly received piece of data might be of no use. In this paper, we aim to incorporate the various data quality metrics into a new sector; data obsolescence. In [10], the authors prove the close relativity between obsolete spatial information and user influence, showing how most advancing technologies are purposed to improve the users' experience and simplify their life. Knowingly, spatial data is one the sectors in need of up-to-date data on a real time basis; otherwise, a potential risk is likely when using map services due to false decision making. That's why the authors of [10] stress on the need of geographical obsolescence identification for keeping data as complete as possible. The latter is of one the fields connected environments encompasses, and the authors agree there exist no strict metrics to evaluate obsolescence. The more the data collected, the more obsolescence is appearing as an issue requiring solutions. This being a reflection of the research gap data obsolescence encounters, in this paper, we investigate data obsolescence metrics and definitions that will help detect data obsolescence within connected environments.

III. Related Work

One might confuse data obsolescence with data quality (DQ) which can explain the few to no research done on the formers' connection. DQ is a widely studied topic described as the wholeness of an entity (data) capable of satisfying the user's needs. In other words, DQ is usually subjective to its intended use, difficult to assess as the quality properties are set by the user. Each property is provided with a collective framework, named data quality dimensions (DQD), to measure data quality and assess its requirements. While DQDs are argued to be domain dependent, standard metrics include: (1) Intrinsic: comparing similar data points to locate any abnormalities, (2) Representational: sampling techniques to view, understand, and manipulate data, (3) Accessibility: where and how the user is accessing the data, and (4) Contextual: driving data based on the data contextual circumstances it belongs to. Table 1 classifies some of the data indicators contained within these four pillars [7] [15] [16] [21].

However, when it comes to data context such as relevance, completeness, added-value, and appropriate data amount, there is still no adequate solution considered [7]. DQDs have been around for quite some time, yet no consensus on a generic methodology is to be found. With the ongoing integration of both worlds, physical and virtual, there are seamless amounts of new systems heavily relying on such data yet minimal focus

TABLE I
STANDARD DATA QUALITY DIMENSIONS

Indicator	Standard Definition
Intrinsic	
Correctness	Free-of-error dimension level where real-world entities, events, or concepts match the intent it is to capture
Accuracy	Comparison level of data in question to a referenced value
Precision	Distortion level created while converting different measurements to one another
Trustworthiness & Credibility	Reputation and authenticity level
Consistency	Coherency/contraction level between the data in question and its previous versions or related comprehensive data
Compliance	Adherence levels to standards, conventions, or regulations
Uniqueness	Redundancy levels within the dataset
Representational	
Understandability	Comprehensivity level of the data in question in means of appropriate languages, symbols, and units
Interoperability	Level to which different systems, technologies, or components can seamlessly work together and exchange information effectively
Ease of operation	Data manipulation level (e.g. updating, moving, aggregating, reproducing, customizing)
Accessibility	
Accessibility	Level to which data are available or easily and quickly retrievable
Security	Restriction level to appropriately to maintain data's security
Traceability	Level to which data are well documented, verifiable, and easily attributed to a source
Availability	Data accessibility level for an intended use when retrieving the data in question is necessary or required
Contextual	
Timeliness	Extent to which data is up-to-date and accurately represents the current state of the phenomenon or process it reflects
Currentness	Percentage of data representing current values; recent data and not from a previous or following form of time
Completeness	Existence of all relevant data to satisfy the user requirement; level of missing data
Relevancy	Level to which data is related to the content at hand
Data amount	Level to which volume of data is sufficient
Value-added	Potential level of data to provide new beneficiary advantages
Cost-effectiveness	How reasonable the cost of collecting data is

on its integrity [13]. While some systems and applications consider some dimensions and leave others out [7], they seem to lack the attention deemed necessary. That's why, this paper, being first in its approach, will adopt and define such crucial dimensions with a new light putting forward methodologies per indicator needed for the detection of data obsolescence in connected environments. In section 3, we delve in to these major indicators' definitions and computations, chosen as the necessities of this obsolescence detection.

Similar confusion is faced by data popularity; a parameter to track how often a piece of data is requested by the system's sites. "Most requested" is not sufficient alone as a metric to measure the data's usefulness especially that data popularity does not investigate how that data is being used by the requestor. "What about the data that are not requested? Are they not useful? What about the data that has just arrived? Might there be missing data?" are a few questions that present a gap. Data popularity is important, but it's also important to know that it's commonly used for data replication management; the stringent conditions needed for data sharing, storing and transferring within distributed systems [14]. Data management concerns, such as eliminating, duplicating, and transferring between disks and caches, are often determined by counting how many times the files have been accessed [3].

IV. PROPOSAL

In this section, we propose formal definitions for data obsolescence and the most significant data quality measures in the field of connected environments.

A. Data obsolescence as an independent terminology

When data becomes insignificant over time and is no longer valid for its intended use, data obsolescence is the phenomenon in depiction. Technological changes are among the top factors causing obsolescence especially after the widespread use of the internet and mobile devices, which has revolutionized the way we communicate and access information. Connected environments, leading this technical revolution, are subject to data obsolescence and require treatment for the data items collected by the sensors. Therefore, in connected environments, we define:

- "Data Obsolescence" as the state wherein data, collected via devices and sensors, is no longer rendered significant, effective, and applicable in the content of its existence.

B. Definitions & preliminaries

As aforementioned, spatial–temporal redundancies face critical lack of treatment with few to none when speaking of connected environments. A number of factors, including modifications in land use, population demography, and infrastructure, can lead to obsolescence especially spatially considering dynamic sensors. This refers to the phenomena wherein information pertaining to a particular place, region, or location, becomes outdated or irrelevant ceasing to serve the original purpose. Even if the sensor was static and not moving, connected environments are subject to spatial obsolescence especially that sensors sense the encompassing surrounding and not necessarily only the data part the system is in need of. Figure 1 visualizes a simple connected environment as a group of zones (e.g. Z0) each having a group of sensors/devices (e.g. D1). Device D2 is a camera monitoring the surrounding locale; however, only space S1 is the data we are interested in. Accordingly, this depicts a spatial obsolescence since a portion of the spatial data the camera collects is outside the relevant space. Taking into consideration that the camera is rotating, at some angle, the whole data might be spatially obsolete with zero relevance. This is only one type and one scenario depicting the need of data treatment, again reflecting the existing research gap.

The aim is not only providing obsolescence with an official definition, but also formulating data estimation definitions and

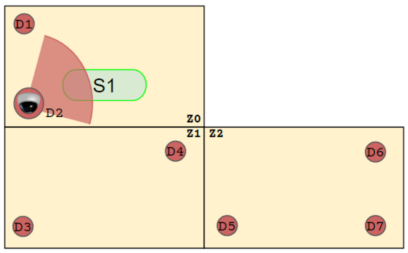

Fig. 1. Sample connected environment

metrics forming the infrastructure necessary for its detection. Us making the first step in comprehending the gap data obsolescence is facing, we focus on connected environments and choose predefined DQDs (Table I). Knowing that such DQDs do not offer computational methodologies, we put forward 19 definitions and demonstrations filling in such gap. Even though the latter are connected environments specific, many are characterized with extension capabilities making them useful for many other technological domains.

Definition 1 (Temporal characteristic.) A temporal characteristic *t* represents when the data was collected:

$$t =< format, v > where: \quad (1)$$

- *format is the type of the temporal stamp (e.g., date, time, datetime, interval, duration)*
- *v is the temporal value expressed in the specified format*

■

Definition 2 (Spatial characteristic.) A spatial characteristic *l* represents information pertaining to a particular place, region, or location:

$$l =< P > where: \quad (2)$$

- *P*=$[p_0, p_1, ..., p_n]$ *is the list of points*
- $p =< longitude(decimal), latitude(decimal), order(integer) >$ *connected based on the specified order to form the required shape (e.g., point, rectangle, square)*

■

Definition 3 (Sensor.) A sensor *s* refers to a physical sensor capable of measuring its feature of interest:

$$s =< type, D, c, a, ca, l, f, fo, sr, se, mse, rrt, mnt, STATUS, m_D > where: \quad (3)$$

- *type (string) refers to the category of the sensor (e.g., temperature, proximity)*
- *D*=$[d_0, d_1, ..., d_n]$ *(Data Item) is the set of the data measured by the sensor saved on an external data storage*
- *c (Gigabyte) is the storage capacity of the aforementioned data storage*
- *ca* $(metric\_unit^2)$ *is the coverage area of the sensor*
- *l is the spatial property specifying the location of the sensor* ***(cf. Def 2)***
- *f (Hertz) is the frequency of sensing*
- *fo is the format of the measured data*
- *sr* = $< min, max >$ *is the sensor range (decimal values)*
- *se (percentage)* = $1 - \frac{sensor\_sensitivity\_in\_documentation}{sr.max}$ *is the sensitivity of the sensor*
- *rp (percentage)* = $\frac{actual\_response\_time}{recommended\_response\_time}$ *is the responsiveness of the sensor*
- *mnt is the manufacture response time of the sensor*
- *STATUS* = $[st_1, st_2, ..., st_i]$ *where* $st_i =< state_i, t_i >$ *is the state (e.g., stopped, active, failed) of the sensor at the respective datetime* ***(cf. Def 1)***
- m_D *(set) is the sensor quality metrics including:*
  - *a (percentage) is the margin of error of the sensor*
  - *av (percentage) is the availability of the sensor*
  - *cr (percentage) is the credibility of the sensor*
  - *ct (percentage) is the contradiction of the sensor*
  - *cn (percentage) is the correctness of the sensor*
  - *fc (percentage) is the format consistency of the sensor*

■

Definition 4 (Zone.) A zone *z* designates is a physical geo-location capable of hosting sensors:

$$z =< type, l, S, a_z > where: \quad (4)$$

- *type (string) specifies whether the zone is a covered - with at least one sensor - or uncovered - with no sensors - area*
- *l is a spatial characteristic representing the particular shape of a zone* ***(cf. Def 2)***
- *S*=$[s_0, s_1, ..., s_n]$ *is the set of sensors* ***(cf. Def 3)*** *deployed in z*
- a_z *(percentage) is the inference of the zone reflecting the impact on the sensors by its environmental properties*

■

Definition 5 (Environment.) An environment *E* refers to a physical infrastructure composed of one or multiple zones:

$$E =< Z > where: \quad (5)$$

- *Z*=$[z_0, z_1, ..., z_n]$ *is the set of zones z* ***(cf. Def 4)***

■

Definition 6 (Individual Data.) A data *d* represents the measurement captured by a sensor in a particular time and location:

$$d =< s\_id, v, t, l, an, m_d > where: \quad (6)$$

- *s_id (string) is the sensor id (e.g. temperature_sensor_1)*
- *v (decimal) is the data value the sensor measured (e.g., 21 °C)*
- *t the temporal property representing when the data was sensed* ***(cf. Def 1)***
- *l the spatial property representing where the data was sensed* ***(cf. Def 2)***
- *an (integer) is the number of times the individual data was accessed or requested*
- m_d *(set) is the single data's quality metrics including:*
  - *tl (percentage) is the timeliness of the data*
  - *sa (percentage) is the spatial accuracy of the data*
  - *ac (percentage) is the accessibility of the data*
  - *pl (percentage) is the popularity of the data*
  - *rl (percentage) is the relevance of the data*

■

Definition 7 (Query.) *q* is a query received by the connected environment. It is related to three properties:

$$q =< q\_id, q\_t, q\_i > where: \quad (7)$$

- *q_id is the query id*
- *q_t is the time at which the query is sent for execution*
- *q_i* = $< k, z, t >$ *incorporates parameters that identify the query's fields of interest:*

– *k is the attribute being searched*
– *z is the data's location being searched* ***(cf. Def 2)***
– *t is the time of interest being searched* ***(cf. Def 1)*** ■

Definition 8 (Coverage area.) A coverage area *ca* designates the spatial limit beyond which the sensor is not capable of catching values:

$$ca =< l, z > where: \quad (8)$$

- *l represents the spatial property of the coverage area limited by the boundaries of z* ***(cf. Defs 4, 5)*** ■

Definition 9 (Accuracy.) Accuracy *a* resembles the margin of error in the measurement of the sensor:

$$a(s) = a_E + a_{fixed} \; where: \quad (9)$$

- a_{fixed} *is the fixed accuracy (percentage) provided by the sensor's documentation (e.g.,* $\pm 2\%$*)*
- a_z *is the additional accuracy (percentage) defect reflecting the impact of z* ***(cf. Def 4)*** *on the sensor* ■

Definition 10 (Availability.) Availability *av* represents the percentage of time the sensor was available and successfully responding to the queries:

$$av(D) = \frac{\sum_{i=0}^{|D|} d_i.an}{\mid Q_D \mid} \; where: \quad (10)$$

- $Q_D = \{q \in Q \mid q \text{ has } S \text{ as the source of interest}\}$ *is a list of all the queries that request the data item D* ■

Definition 11 (Credibility.) Credibility *cr* provides the extent to which the sensor can be trusted with the measurements it is providing:

$$cr(D) = K_1 * rt + K_2 * rp + K_3 * se \mid \sum_{i=0}^{n} K_i = 1 \quad (11)$$

(cf. Def 3) *where:*

- $rt = 1 - \frac{|\{d \in D | d == sr.max\}| + |\{d \in D | d == sr.min\}|}{|D|}$ *is the resolution of the sensor resembling the percentage of D having the min and max saturation data values* ■

Definition 12 (Contradiction.) Contradiction *ct* is the discrepancy present within the data item reflecting the percentage data drift its encountering:

Algorithm 1: Contradiction Computation Via Data Drift Check

```
Input    : D
Output   : ct
Variables: drift_percent, i
drift_percent = 0;
for i = 2 to i = size(D) do
    /* Equality_function infers similarity between two consecutive data
       individuals                                                    */
    drift_percent += Equality_function(d_i, d_{i-1}));
ct = drift_percent / size(D);
return ct;
```

■

Definition 13 (Correctness.) Correctness *cn* provides the extent to which the sensor's measurements are affected by external conditions:

$$cn(S) = K_1 * sf(S) + K_2 * frf(S) \mid \sum_{i=0}^{n} K_i = 1 \; where: \quad (12)$$

- $sf(D) = 1 - \frac{off\_time}{on\_time}$
$= 1 - \frac{\sum_{i=1}^{n}(STATUS_{(i)}.t - STATUS_{(i-1)}.t) \cdot \mathbb{I}(\text{STATUS}_{(i)}.\text{state}=\text{active})}{\sum_{i=1}^{n}(STATUS_{(i)}.t - STATUS_{(i-1)}.t) \cdot \mathbb{I}(\text{STATUS}_{(i)}.\text{state}\neq\text{active})}$
is the percentage of failures/stoppages the sensor faces according to its status logs
- $frf = 1 - \frac{f}{recommended\_freq}$ is the frequency with respect to the recommended one ■

Definition 14 (Format consistency.) Format Consistency *fc* checks how compatible and the change of compatibility in the sensor's data format:

$$fc(D) = 1 - \frac{\mid D_{fd} \mid}{\mid D \mid} where: \quad (13)$$

- $D_{fd} = \{d.f \in D \mid d.f \text{ appears only once in } D\}$ *is a list of the distinct formats that exists in the data item D* ■

Definition 15 (Timeliness.) Timeliness *tl* is the degree to which data has attributes that are of the right age in a specific context of use with respect to the queries:

Algorithm 2: Timeliness Calculation

```
Input    : d, Q
Output   : avg_t_s
Variables: T_S
avg_t_s ← 0;
T_S ← new_list();
/* allen_temporal detects relation of two temporal intervals (e.g., equal,
   intersect)                                                          */
for q in Q do
    if allen_temporal(q.q_i.t, d.t) ∈ ["equal", "include"] then
        T_S ← 1;
    else if allen_temporal(q.q_i.t, d.t) == "outside" then
        T_S ← 0;
    else
        T_S ← (UNION(q.q_i.t, d.t)/INTERSECTION(q.q_i.t, d.t));
avg_t_s = SUM(T_S) / COUNT(T_S);
return avg_t_s;
```

■

Definition 16 (Spatial accuracy.) Spatial accuracy *sa* is the degree to which data has attributes that are of the right location in a specific context of use with respect to the queries:

Algorithm 3: Spatial Accuracy Calculation

```
Input    : d, Q
Output   : avg_s_s
Variables: S_S
avg_s_s ← 0;
S_S ← new_list();
/* DE_9IM detects relation of two spatial objects (e.g., equals, contains,
   disjpoint, intersects, touches, crosses, within, overlaps)          */
for q in Q do
    if DE_9IM(q.q_i.z, d.l) ∈ ["equals", "contains"] then
        S_S ← 1;
    else if DE_9IM(q.q_i.z, d.l) ∈ ["disjoint", "touches"] then
        S_S ← 0;
    else
        S_S ← (UNION(q.q_i.z, d.l)/INTERSECTION(q.q_i.z, d.l));
avg_s_s = SUM(S_S) / COUNT(S_S);
return avg_s_s;
```

■

Definition 17 (Accessibility.) Accessibility *ac* represents the storage means of the data and how applicable it is to access:

$$ac(d) = \frac{d.an}{\mid Q_d \mid} \; where: \quad (14)$$

- $Q_d = \{q \in Q \mid q \text{ has } d \text{ as the data of interest}\}$ *is a list of all the queries that request the individual data d* ■

Definition 18 (Popularity.) Popularity *pl* is the level of interest or attention that an individual data receives with respect to the data item it belongs to:

$$pl(d) = \frac{d.an}{\sum_{i=0}^{|D|} d_i.an} \quad (15)$$

■

Definition 19 (Relevancy.) Relevancy *rl* is the degree to which data is within the valid expected range:

$$\text{Relvancy\_function}^{\theta}(d) = \begin{cases} 1, & d \in valid\_range \\ 0, & d \notin valid\_range \\ similarity & valid\_range \in \phi \end{cases} \quad (16)$$

- *similarity (e.g., cosine) computes the degree to which data d is close to the respective data item and its logs*

■

As noticed from Table 1, although data quality is widely investigated, it's usually covered within theoretical aspects. Data quality dimensions are refereed back to one of the four main categories aforementioned; namely, intrinsic, representational, accessibility, and contextual. In other words, when coming across a certain indicator (e.g. consistency, contradiction), we go up the hierarchy checking whether this indicator focuses on the internal consistency, data viewing, ease of access, or data significance when the focus should be more on the indicator itself and how to compute it. Detecting this gap, we create clear definitions and specifications per indicator providing the hierarchy with its infrastructure. In other words, we are providing the first step for the data quality indicators to be as independent while giving space and enlightenment for expansion towards other domains. Consequently, while our definitions and metrics are inspired by connected environments, they are applicable to many other technological domains. As an end result, all indicators work collectively to provide the connected environments with what's needed as an initial step for obsolescence detection.

V. Future Scope

This paper investigates DQDs made helpful for data obsolescence detection within connected environments. Each indicator tackled is provided with a computation procedure producing a numerical output. However, how all these estimation will be grouped together in order to produce a single obsolescence percentage will be the role of future work. Based on the connected environment properties, each dimension should be assigned a specific weight describing its contribution to the overall obsolescence detection. Additionally, this paper shows the initial contribution towards practical DQDs quantifications; however, it is specific to connected environments. Accordingly, the future scope also includes what and how dimensions can be extended towards other domains.

VI. Conclusion

In the age of rapid technology, the emphasis is more on data security and interoperability with less emphasis on data integrity; namely, Data Obsolescence. Depending on the context in which it is used, data obsolescence might exist in practically any area of study and have different definitions. In connected environments, we define "Data Obsolescence" as the state wherein data, collected via devices and sensors, is no longer rendered significant in the content of its existence. In this context, we propose and characterize major data quality computational indicators offering data obsolescence the baseline needed for its detection. We put forward 19 definitions resembling metrics within the connected environment with extension capability towards other technological domains. The paper examines how each indicator is calculated, as to how they work together to provide a single obsolescence estimation, that will be within the paper's future scope.

References

[1] Alieksieiev, V., Ivasyk, G., Pabyrivskyi, V., Pabyrivska, N.: Big data aggregation algorithm for storing obsolete data (03 2018)

[2] Bashir, O.: Managing obsolescence in information technology. In: Working document presented in the National IT conference (2000)

[3] Beermann, T., Chuchuk, O., Girolamo, A.D., Grigorieva, M., Klimentov, A., Lassnig, M., Schulz, M., Sciaba, A., Tretyakov, E.: Methods of data popularity evaluation in the ATLAS experiment at the LHC. EPJ Web of Conferences **251**, 02013 (2021)

[4] Bouzeghoub, M.: A framework for analysis of data freshness. In: Proceedings of the 2004 international workshop on Information quality in information systems. ACM (Jun 2004)

[5] Bradley, M., Dawson, R.J.: An analysis of obsolescence risk in it systems. In: Software Quality Management VI, pp. 209–217. Springer London (1998)

[6] Bradley, M., Dawson, R.: An analysis of obsolescence risk in it systems. Software Quality Control **7**, 123–130 (07 1998). https://doi.org/10.1023/A:1008808708860

[7] Byabazaire, J., O'Hare, G., Delaney, D.: Data quality and trust: Review of challenges and opportunities for data sharing in iot. Electronics **9**(12), 2083 (2020)

[8] Chaieb, S., Hnich, B., Mrad, A.B.: Data obsolescence detection in the light of newly acquired valid observations. Applied Intelligence **52**(14), 16532–16554 (Mar 2022)

[9] Finger, M., Da Silva, F.: Temporal data obsolescence: modelling problems. In: Proceedings. Fifth International Workshop on Temporal Representation and Reasoning (Cat. No. 98EX157). pp. 45–50. IEEE (1998)

[10] Glushkov, A., Belyakov, S., Belyakova, M.: Intellectual obsolescence detection method of spatial data using historical data. In: 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT). IEEE (Sep 2017)

[11] Habermann, A.: Automatic deletion of obsolete information. Journal of Systems and Software **5**(2), 145–154 (May 1985)

[12] Junge, I., van der Velden, M.: Obsolescence in information and communication technology: A critical discourse analysis. In: This Changes Everything – ICT and Climate Change: What Can We Do?, pp. 188–201. Springer International Publishing (2018)

[13] Liu, C., Nitschke, P., Williams, S.P., Zowghi, D.: Data quality and the internet of things. Computing **102**(2), 573–599 (2020)

[14] Ma, J., Liu, W., Glatard, T.: A classification of file placement and replication methods on grids. Future Generation Computer Systems **29**(6), 1395–1406 (Aug 2013)

[15] Makhoul, N.: Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring. Advances in Bridge Engineering **3**(1) (2022). https://doi.org/10.1186/s43251-022-00068-9

[16] Mansouri, T., Sadeghi Moghadam, M.R., Monshizadeh, F., Zareravasan, A.: Iot data quality issues and potential solutions: A literature review. The Computer Journal **66**(3), 615–625 (2021). https://doi.org/10.1093/comjnl/bxab183

[17] Noueihed, H., Harb, H., Tekli, J.: Knowledge-based virtual outdoor weather event simulator using unity 3d. The Journal of Supercomputing **78**(8), 10620–10655 (Jan 2022)

[18] Rojo, F.J.R., Roy, R., Shehab, E.: Obsolescence management for long-life contracts: state of the art and future trends. The International Journal of Advanced Manufacturing Technology **49**(9-12), 1235–1250 (Dec 2009)

[19] Romero Rojo, F.J., Roy, R., Shehab, E.: Obsolescence management for long-life contracts: state of the art and future trends. The International Journal of Advanced Manufacturing Technology **49**, 1235–1250 (2010)

[20] Schulze, F.A., Arndt, H.K., Feuersenger, H.: Obsolescence as a future key challenge for data centers. In: Progress in IS, pp. 67–78. Springer International Publishing (Dec 2020)

[21] Teh, H.Y., Kempa-Liehr, A.W., Wang, K.I.K.: Sensor data quality: A systematic review. Journal of Big Data **7**(1), 1–49 (2020)