An Approach for Assessing Industrial IoT Data Sources to Determine their Data Trustworthiness

Abstract

Trustworthy data in the Industrial Internet of Things are paramount to ensure correct strategic decision-making and accurate actions on the shop floor. However, the enormous amount of industrial data generated by a variety of different sources (e.g. machines, sensors) is often of poor quality (e.g. unreliable sensor readings). Research suggests that certain characteristics of data sources (e.g. battery-powered power supply, wireless communication) contribute to this poor data quality. Nonetheless, to date, much of the research on data trustworthiness only focused on the data values to determine the trustworthiness. Consequently, we propose to pay more attention to the characteristics of data sources in the context of data trustworthiness. Thus, this article presents an approach to assess Industrial Internet of Things data sources to determine their data trustworthiness. The approach is based on a meta-model decomposing data sources into data stores (e.g. databases) and providers (e.g. sensors). Further, the approach provides a quality model comprising quality-related characteristics of data stores to determine their data trustworthiness. Moreover, a catalog containing properties of data providers is presented to infer the trustworthiness of their provided data. An industrial case study revealed a moderate correlation between the data source assessments of the proposed approach and experts.

Keywords: Data trustworthiness, Data source assessment, Industrial Internet of Things

1. Introduction

First presented by General Electric in 2012 [1], the general concept of the Industrial Internet of Things (IIoT) has become an essential component of current efforts (e.g. [2, 3]) to digitally transform various industries around the world. The general concept of the IIoT refers to integrating and connecting all physical devices (e.g. machines, sensors, actuators) and industrial control systems (ICSs) with traditional information systems and business processes [4]. To enable this integration of diverse industrial assets, the IIoT uses several modern technologies such as cloud computing, radio frequency identification (RFID), cyber-physical systems (CPSs), wireless sensor networks (WSNs) and mobile technologies (e.g. 5G) [5]. The resulting system of interconnected industrial assets generates a large amount of industrial data which can be used in several ways to optimise industrial operations or to generate new innovative business models [5]. Reduced resource consumption, minimized costs, improved product quality or increased productivity are just some of the many benefits the IIoT offers [6, 7]. To realize these profits, the IIoT is used across several industries, such as transportation, energy, telecommunications, agriculture and manufacturing [6]. Together, these industries will generate a global IIoT market size of more than $900 billion by 2025 [8].

At the heart of the IIoT are advanced analytical systems that provide valuable insights and enable intelligent operations based on the gathered data [9]. Classical examples are predictive maintenance techniques where machine data are analyzed to predict component failures in order to avoid machine downtime [10]. More complex examples are self-organizing networks of devices that use compound machine learning algorithms to adjust themselves completely without human intervention [4].

However, these analytical systems rely heavily on the quality of the data provided. Thus, low quality data (e.g. dropped or unreliable sensor readings) can lead to wrong decisions or even cause misleading actions on the shop floor that harm humans. As a proxy measure for data quality [11, 12], data trustworthiness therefore gained significant interest in the IIoT in the last few years [13, 14]. In broad terms, data trustworthiness refers to the probability that the data provided by data sources are correct [15, 16]. To determine this probability, research to date (e.g. [17, 18, 19]) has tended to focus on the extension of the data (i.e data

values) rather than on the sources which provide these data. The few studies that take data sources into account, in turn, only assess them on the basis of data values (e.g. [20, 15]), completely ignoring their intrinsic characteristics (e.g. data schema, hardware/resource constraints, environmental influences).

Nonetheless, it is precisely these characteristics that are proven to be a significant factor influencing the quality of the data provided by data sources [15, 21, 20, 16, 22]. In fact, data quality in IIoT systems is often influenced by common intrinsic technical limitations of IIoT devices such as energy, memory and computing constraints [23, 24]. Further, environmental factors such as harsh conditions or damage can influence the quality of industrial data transmitted from sources. In addition, the widespread wireless communication used by various IIoT devices tends to be more error-prone than traditional wired communication [25, 26, 27]. Accordingly, we argue that it is a necessity to consider the characteristics of IIoT data sources to determine their data trustworthiness.

To address this need, this paper deals with assessing IIoT data sources based on their characteristics to infer the trustworthiness of their provided data. In particular, a *data source assessment approach* is presented that determines the probability that IIoT data sources provide correct data in terms of data accuracy, data completeness, data consistency, data credibility and data currentness. The presented approach consists of a meta-model, a quality model and a catalogue. In more detail, the contributions of this work are the following:

- A *meta-model* that enables the representation of IIoT data sources by means of data stores (e.g. databases, data lakes) and data providers (e.g. sensors, scanners).
- A *quality model* comprising characteristics related to the intrinsic quality of data stores (e.g. availability, metadata quality) to determine their data trustworthiness.
- A *catalogue* based on properties of data providers (e.g. battery-powered, mobility) to infer the trustworthiness of their provided data.
- An *evaluation* of the data source assessment approach in the context of an industrial case study. The results of the evaluation indicate that the assessments conducted with the approach can be treated as valid.

The remaining article specifically focuses on the manufacturing sector of the IIoT and is structured as following. Section 2 first provides relevant background information on the IIoT and data quality. Afterwards related work on data trustworthiness and assessing data sources is described. In Section 3, the developed data source assessment approach is presented. Following, Section 4 describes the empirical validation of the developed approach. Limitations and implications of the results are discussed in Section 5. Finally, Section 6 concludes the paper and outlines future work.

2. Background and Related Work

This section introduces necessary background information on the IIoT (Section 2.1) and data quality (Section 2.2). Afterwards, a brief overview about related work in the context of data trustworthiness and assessing data sources is given in Section 2.3.

2.1. Industrial Internet of Things

In contrast to the general IoT which is human centered, focusing on rather low-priced, up-to-date smart consumer electronic devices, the IIoT aims to connect all industrial assets such as sensors, machinery and ICSs with each other and beyond with back-end information systems and the Internet [4]. Through the usage of modern technology and complex software stacks, the IIoT enables harvesting an enormous amount of data that was previously often impossible or unprofitable to collect [28]. For example, wireless sensors attached to physical devices or placed at any location at the shop floor generate valuable new data that can be send directly to the cloud [29]. Thus, a lot of heterogeneous physical devices (e.g. machines, robots, pumps, engines) and their corresponding tooling equipment (e.g. cutting equipment, gauges, dies) are acting

as data sources within the IIoT. In addition, manufacturing information systems (e.g. MES, ERP, SCM), products or further devices (e.g. RFID readers, video cameras) often serve as sources for the data collected.

The tremendous amount of available data can be used for various innovative and intelligent applications. Possible applications span from rather simple monitoring scenarios that provide increased knowledge about the physical process to very complex scenarios in which physical devices build self-organizing systems (e.g. CPSs) that are driven by complex algorithms (e.g. machine learning) [30, 31]. Two of the currently most prominent application scenarios [32] are predictive maintenance and product quality assurance. The aim of predictive maintenance is to predict when the next failure of a machine would take place and conduct maintenance before this event actually happens [33, 34]. Product quality assurance applications generally aim to increase the quality of the product to be manufactured or to improve the quality of the entire production process. Through processing a variety of different data (e.g. historical data, machine condition data, process-related or product-specific data), machine learning techniques are able to predict a product's final quality [35], detect root causes of defects as well as to detect abnormal states of a machine or the onset of machine components degradation.

Before the data can actually be used for such applications, it has to pass through different networks with varying protocols. Following, a widely accepted architectural pattern is used to describe the typical data flow within an IIoT system. Namely, the three-tier architecture pattern, is a simplified and abstracted view of an IIoT system and comprises the edge, platform and enterprise tier [36]. Each tier plays a specific role in processing the data flow as shown by Figure 1. Unless otherwise stated, the following description of each tier is mainly based on [36, 4, 37].

Edge tier. The edge tier is responsible for collecting data from all ICSs and field devices (e.g. sensors, actuators, controllers, machines) via local types of networks (e.g. Bluetooth, Wi-Fi, Fieldbus). These networks form the so-called *proximity network* which enables the connectivity between all industrial assets and connects to the IIoT edge gateway. In detail, the concrete flow of the data within the proximity network depends on whether the industrial data producer at the shop floor has smart characteristics (i.e. computational and IP network capabilities) or not.

The data flow of *smart devices* is outlined at the right-hand side of the edge tier in Figure 1. Such devices are able to send their generated data directly to the platform tier by using Internet application protocols such as Message Queuing Telemetry Transport (MQTT) or Hypertext Transfer Protocol Secure (HTTPS). More frequently, they exchange data with the IIoT edge gateway by implementing Open Platform Communications Unified Architecture (OPC UA) functionality.

In contrast to smart devices, the data flow of the conventional industrial equipment (left-hand side of the edge tier in Figure 1) starts with sensors which generate massive industrial data. These sensors are typically installed within or attached to physical equipment. The generated data, typically measurements (e.g. temperatures, pressure), are then transmitted directly via fieldbus networks (e.g. CAN bus, Profibus) to control devices (e.g. DCS, PLC).

In a next step, ICS store the incoming data (e.g. in Historians) and provide user interfaces to monitor and steer the production processes. Moreover, these systems forward the data often to on-premise information systems (e.g. ERP, MES) for further action (e.g. order management, production planning). The data transmission between the ICS and the control devices can either be executed through native industrial control bus protocols (e.g. PROFINET, Modbus) or, more common, using some middleware software.

To enable the IIoT, the IIoT edge gateway forwards the data to the platform tier using the *access network*. This network is usually implemented as a virtual privacy network or mobile network and uses protocols such as MQTT or HTTPS.

Platform Tier. At the platform tier, basic transformations, storage and analysis services for the receiving data are offered. Consequently, the platform tier provides functions to monitor and optimize the systems at the edge tier. Furthermore, this tier also offers management and provisioning functions on industrial devices located at the shop floor. In detail, an IIoT hub is responsible for dispatching the incoming data to the different services or storage systems. The services and the data storage are usually hosted in the cloud or in a remote data center. Moreover, the platform tier also processes and forwards control messages sent from

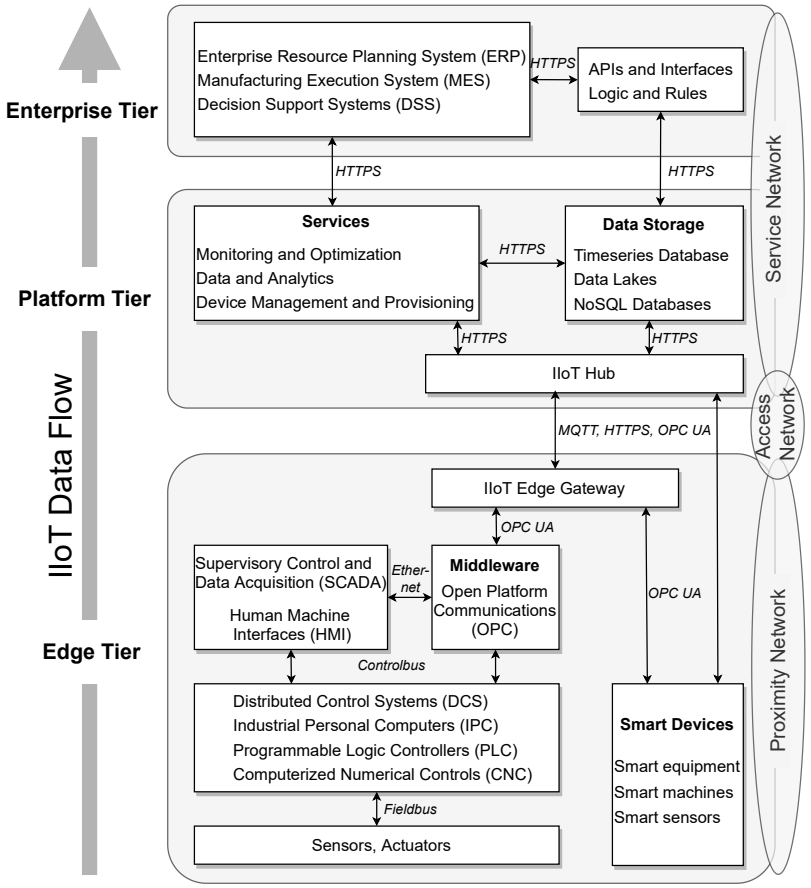

Figure 1: Industrial internet of things data flow.

the enterprise tier to the edge tier. All services of the platform tier are connected to the upper enterprise tier through the *service network* (e.g. virtual privacy network, Internet). The service network can be a virtual privacy network or the public Internet.

Enterprise Tier. The enterprise tier finally receives the processed data from the lower tiers. At this tier, domain-specific applications such as ERP or MES and decision support systems are hosted. Therefore, application and high-level business logic are implemented by the enterprise tier. Based on these logics, control commands are sent to the platform and edge tiers. Further, the enterprise tier provides graphical user interfaces (GUIs) or application programming interfaces (APIs) to end users.

2.2. Data Quality

The constantly increasing amount of generated data and the resulting possibility to make data-driven decisions makes data quality an increasingly important aspect to consider. However, research on data quality already started decades ago [38, 39]. Based on its context-dependence and the different possible perspectives under which data quality can be considered, various data quality frameworks [40], models (e.g. for Web Portals [41], Linked Data [42], Big Data [43]) and assessment methodologies [44, 45] have been proposed. In addition, a considerable number of characteristics (often also named dimensions or attributes) have been proposed that aim to describe desirable aspects of data quality.

Over time, also different definitions of data quality emerged. The most basic and prominent definitions are based on the concept of "fitness for use" [39]. In simple terms, this concept emphasizes the importance of the intended usage of the data. Thus, the quality of data should be considered from a consumer perspective because only the data consumer can judge whether the data are fit for use or not. Based on this concept, Wang & Strong [39] define data quality as "*data that are fit for use by data consumers*".

Throughout this article, we adopt ISO/IEC's [46] understanding of data quality which is close to the concept of "fitness for use". On an abstract level, the standard defines data quality as the degree to which data satisfy defined requirements. The standard further classifies data quality characteristics into two main categories, namely *system-dependent* data quality and *inherent* data quality. The latter category comprises quality characteristics that refer to the data itself (i.e. to their values) such as *accuracy*, *completeness*, *consistency*, *credibility* and *currentness*. On the other hand, the system-dependent category describes data quality as degree to which it is reached and preserved within computer systems, thus considering the technological domain in which data are used. Characteristics related to this category are *availability*, *portability* and *recoverability*. The standard also proposes characteristics that are assigned to both categories, namely: *accessibility*, *compliance*, *confidentiality*, *efficiency*, *precision*, *traceability* and *understandability*. An overview about all quality characteristics including their definitions is provided in Appendix Appendix A.

Although research on data quality is quite mature and extensive, the emergence of the IoT paradigm created new substantial data quality challenges (e.g. diversity of data sources, increasing data volume and generation speed). Moreover, also modern data-intensive software systems [47] pose several challenges on ensuring data quality due to their incorporated machine learning algorithms (e.g. entanglement of model, parameters and data sets). For a detailed overview about current data quality challenges see [48, 49, 50, 23].

2.3. Related Work

Data trustworthiness can broadly be described as the probability that the data provided by data sources are correct. According to Rahman et al. [13], 'correct' in this sense refers to the freedom of errors and the up-to-dateness of the data as well as to the fact that the data sources providing the data are reputable. Based on this description of data trustworthiness, there are a variety of related research areas trying to actively contribute to ensuring the correctness of data such as research on data quality in general including data quality assessment and assurance as well as data validation, data filtering or data provenance. The remaining section, however, is limited to related work on data trustworthiness and data source assessment.

2.3.1. Data Trustworthiness

As was pointed out in the introduction to this article, the existing literature on data trustworthiness pays particular attention to the data items themselves rather than on the sources which provide these data. In the following, we describe two articles that take the data source into account in their approaches. However, the computation of the trustworthiness of the data provided by the data source is again based on data values and not on data source characteristics.

The first article was published by Dai et al. [20] in 2008. In their work, the authors present a data provenance framework to determine the trustworthiness of data and data providers. Their trust model considers the aspects data similarity, data conflict, path similarity and data deduction to determine the trustworthiness. Although the approach initially use unspecified criteria of data providers to determine their trustworthiness, it is recomputed based on the average trustworthiness of the data provided.

The second approach was proposed by Tang et al. [15]. They present a framework to identify trustworthy sensor alarms within CPS. To improve their proposed trustworthiness inference, the authors use the reliability of a sensor. This sensor reliability is close to our understanding of the data trustworthiness of a data source. However, the computation of the sensor reliability is done on data items only. A comprehensive overview about further data trustworthiness approaches is provided by Bertino [21].

It is worth noting that the interest in the trustworthiness of data within the field of IIoT has grown significantly in recent years. Possible explanations for this may be the widespread use of the IIoT, the high data dependency of IIoT systems or the high demands of industrial environments on such systems (e.g. high fault tolerance, high robustness and safety). A rough overview about current developments and research

trends in this area is provided by Rahman et al. [13]. Further studies that make contributions towards data trustworthiness in IIoT environments are [19, 51, 18]. Nevertheless, these publications again focus only on the data and not on the characteristics of the sources they come from to determine the trustworthiness.

2.3.2. Data Source Assessment

In addition to research on data trustworthiness itself, there are some studies available that deal directly with the assessment of data sources based on their intrinsic characteristics. These studies tend to focus on the domain of administrative (i.e. statistical) or open data and are therefore only applicable to the IIoT to a limited extend. Following, we briefly describe some related work within this context.

Daas et al. [52] developed a framework to assess administrative data sources. In their work, the authors use various intrinsic characteristics of data sources (e.g. punctuality, format, security) to describe the source hyperdimension of their framework. Later, Dufty et al. [53] extended this framework to encompass big data quality using several intrinsic data source characteristics (e.g. complexity, reliability).

A further work to be mentioned is that of Milošević et al. [54] who describe a methodology to assess data sources from different perspectives. In addition to inherent data quality characteristics, the proposed methodology also includes characteristics related to intrinsic characteristics of data sources (e.g. durability, access, complexity).

Further, important characteristics of data sources (e.g. data availability, data format, existence of metadata) are incorporated within the technological perspective of the open data maturity model proposed by Solar et al. [55]. In their article, the authors outline a maturity model as reference for public agencies in implementing open data principles.

3. IIoT Data Source Assessment Approach

In this section, we present an approach to assess IIoT data sources regarding their probability of providing correct data (i.e. their data trustworthiness). The key idea of the approach is to represent data sources in terms of two concepts, namely data stores and data providers. Basically, data providers are assumed to generate data and send them to data stores for storage. However, to assess data sources in terms of their data trustworthiness, criteria for both data stores and data providers are required. Therefore, the approach utilizes quality characteristics of data stores and properties of data providers to determine the trustworthiness of the data they provide.

The remaining section is structured as follows. First, Section 3.1 introduces the underlying *meta-model* of the approach and details the relationship of the different concepts. Afterwards, a *data store quality model* tailored to infer the trustworthiness of data stored in data stores is presented in Section 3.2. Subsequently, Section 3.3 outlines a *catalogue comprising properties of data providers* to determine their probability of providing correct data. Section 3.4 elaborates on the *calculation of the data trustworthiness*. In Section 3.5, a *procedure for applying the assessment approach* is outlined. Finally, an *exemplary application* of the approach is presented in Section 3.6.

3.1. Meta-Model of the Approach

To enable a consistent description of diverse IIoT data sources, we introduce the meta-model (based on UML notation) shown in Figure 2. As reported above, the key concepts of our approach are data stores and data providers. They are each represented by the corresponding meta-class *Data Store* and *Data Provider*. A data provider is defined as a physical entity that generates data, is located at the shop floor and typically does not provide extensive data storage capabilities. Examples of data providers are sensors, scanners, conventional production machinery or industrial equipment. In contrast, a data store can be described as a typical data storage entity that provides comprehensive capabilities of storing and managing data. Data store elements can be flat files, databases or data lakes.

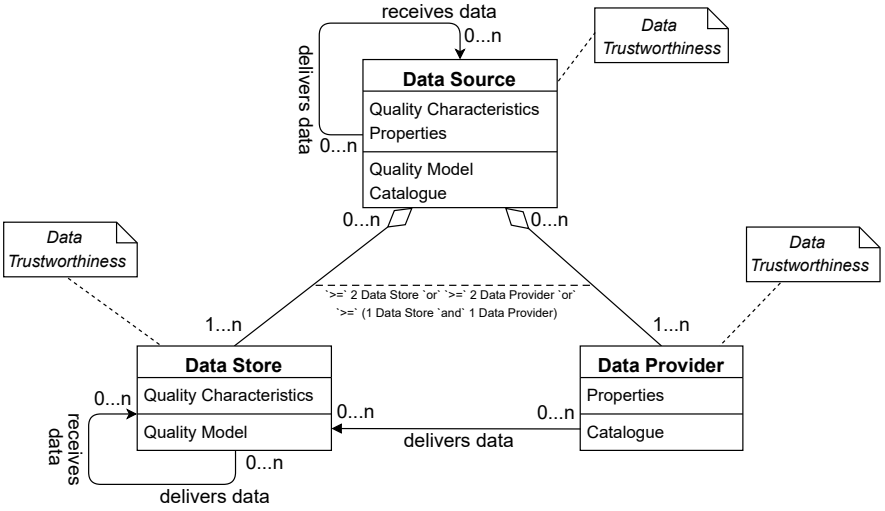

Figure 2: Data source meta-model (UML notation).

The basic idea in our approach is that a data provider delivers its generated data to a data store. This is represented by the association *delivers data* between the data store and data provider classes. In certain cases (e.g. data streaming), however, data providers may deliver data for immediate use without prior storage (e.g. real-time analytics). Further, it is worth noting that a data provider cannot deliver data to another data provider. In contrast, data stores may deliver data to other data stores (recursive association *receives data/delivers data on meta-class data store*).

However, industrial ecosystems usually contain a huge number of different data providers and data stores. For example, production machinery can integrate hundredths of sensors and IIoT networks can include plenty of different data stores. Depending on the concrete scenario (e.g. WSN with thousands of sensor nodes) it would be impossible to consider each single sensor (i.e. data provider) or data store within an industrial ecosystem. Moreover, beside using several sensors, smart machines can either compromise local data stores or are seamlessly connected to networked data storage systems. Resulting, it would be very difficult or time-consuming to consider the individual data stores and data providers of already one smart machine.

To address this issue, we further introduce the concept of a data source (meta-class *Data Source*). The concept of a data source allows to aggregate several individual data stores and data providers. Therefore, a data source can contain several data providers and data stores (represented by two *aggregation relationships*). In detail, a data source can either contain at least one data store and one data provider *or* at least two data stores or data providers. As long as data stores are part of data sources, data sources may deliver data to other data sources (recursive association *receives data/delivers data on meta-class data source*).

A concrete instantiation of the meta-model is shown in Figure 3. The upper part of the figure shows three possibilities to model a scenario where two robot arms deliver data to a database. The first possibility (1) is to model the two robot arms as independent data providers and the database as data store. In case the two robot arms are identical, they can be modeled as one instance of a data source (2). As a last possibility, the database and both robot arms are combined and modeled as data source (3). The lower part of the figure shows a robot arm including an embedded database and some equipped sensors. This setting can be modeled in two ways. First, all sensors are modeled as data providers while the database is modeled as a data store (4). In contrast, the lower right side of the figure illustrates the possibility to model this setting as a single data source (5). However, which option is chosen depends on the concrete application scenario. For example, how many different data sources need to be considered in total or what the concrete objective is.

So far, we introduced the main concepts (i.e. data store, data provider and data source) and its relation-

ships (i.e. delivers/receives data and aggregation). To determine their corresponding data trustworthiness, further concepts need to be introduced. Therefore, the approach introduces both the concept of quality characteristics of data stores and the concept of properties of data providers.

In detail, we propose to determine the probability that a data store provides correct data (i.e. its data trustworthiness) by considering its intrinsic quality characteristics (attribute *Quality Characteristics*). Based on our previous work [56] and further literature (e.g. [22]), we assume that a data store delivers data of high quality when certain of its quality-related characteristics (e.g. availability, data schema completeness) are high and vice versa. To assess these characteristics, we introduce the concept of a quality model (operation *Quality Model*). Thus, the data trustworthiness of a data store (comment notation *Data Trustworthiness*) is determined by assessing its intrinsic quality-related characteristics based on a quality model.

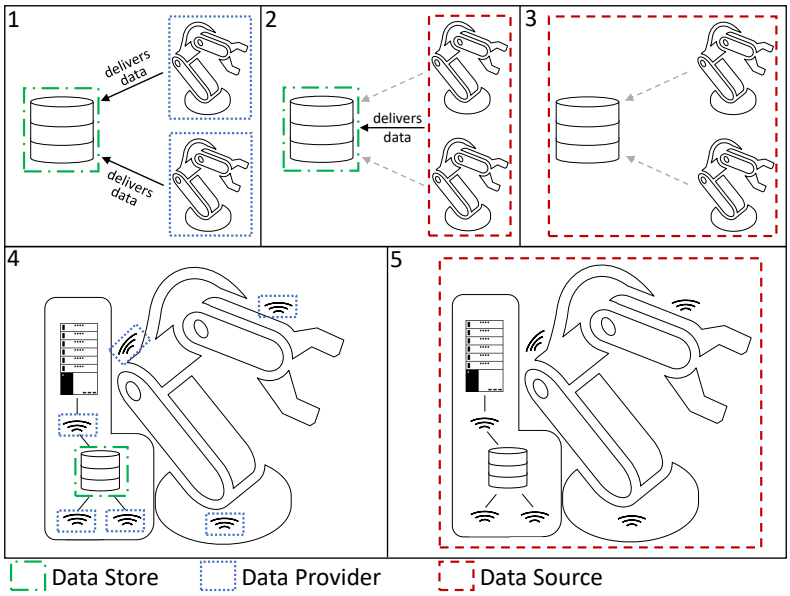

Figure 3: Meta-model instances.

The probability that a data provider delivers correct data is determined by using certain of its intrinsic properties (attribute *Properties*). This idea is based on literature (e.g. [23]) which claims that specific properties of data providers (e.g. mobility, hardware constraints) are likely to influence the quality of the data provided. To assess these properties, we introduce the concept of a catalogue (operation *Catalogue*). Hence, the data trustworthiness of a data provider (comment notation *Data Trustworthiness*) is determined by assessing its intrinsic properties based on a catalogue.

Due to the fact that a data source can contain both data stores and data providers, its data trustworthiness is determined by assessing both quality characteristics and properties. More details on the trustworthiness determination are given in Section 3.4. The next two sections present the development of the data store quality model and the data provider property catalogue.

3.2. Data Store Quality Model

The different steps of the quality model development process are outlined in Figure 4. To determine the quality characteristics of data stores, our first intention was to consider the ISO/IEC 25012 data quality model (see Section 2.2) as it introduces characteristics related to the quality of data stores in its category *system-dependent* data quality. However, this category only provides three characteristics (i.e. availability, portability, recoverability) without any further sub-characteristics. Thus, they were not sufficient for developing the quality model. Therefore, our first step was to conduct a systematic literature review to get a comprehensive initial overview about quality characteristics of data stores. Afterwards, we determined

which of those characteristics are relevant for further consideration (i.e. may influence the quality of the data provided). Hence, each characteristic was assessed in terms of its influence on the five most common *inherent* data quality characteristics (see Section 2.2) and on typical data engineering activities. The resulting set of characteristics was then subdivided into further sub-quality characteristics and properties. Lastly, a checklist was developed to operationalize the quality model. The following sections (given in Figure 4 in parenthesis) describe each step in more detail.

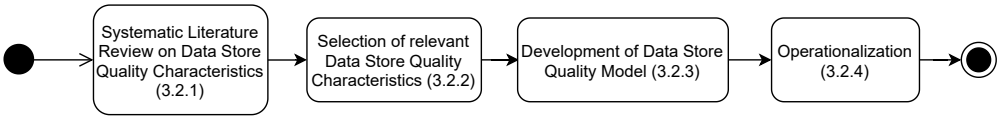

Figure 4: Data store quality model development process.

3.2.1. Literature Review

We followed the guidelines presented by Kitchenham & Charters [57] and Wohlin [58] for conducting our review. As depicted in Figure 5, the process of conducting the review consisted of three main stages compromising different steps which are described below.

Search Process and Source Selection. The review was performed in November and December 2019. As a first step, we conducted a trial search in Google Scholar to identify relevant key words and to determine an appropriate search strategy. Based on this trial search, we identified a set of key words which we combined into nine search string (enclosed in square brackets) as follows: *[('data source' OR 'data store' OR 'information source') AND 'quality model'] OR [('data quality' OR 'schema quality' OR 'data source quality' OR 'information source quality') AND 'intensional'] OR [('data' OR 'information') AND 'source quality'] OR ['intensional data quality' OR 'system-dependent data quality'] OR ['data service quality' OR 'data provision quality'] OR ['data source assessment'] OR ['schema quality'] OR ['data source quality'] OR ['information source quality'].*

Table 1: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Accessible in full text	Non-english articles
Published between 1995 and 2019	Grey and white literature
Addressing quality characteristics related to data stores	Addressing only inherent data quality
	No quality characteristic definitions/descriptions provided
	Too specific (e.g. focus only on administration/linked data sources)

These search strings were then used to perform an initial search on Google Scholar's database. Based on scanning the title, abstract and conclusion of each hit we selected a first set of potential relevant sources. However, due to their generic nature, some key words (i.e. data, data source) returned a very large number of hits, many of which were not relevant to our study. To reduce the number of hits to a manageable size, we restricted the search space by utilizing Google Scholar's Ranking Algorithm. In fact, we assumed that the most relevant articles usually appear at the first few pages. Thus, we only checked the first three result pages of each search string's results (i.e. 30 hits) and only continued if a relevant hit was found at the last page. In total, we excluded 383 articles from an initial pool of 460 scanned sources resulting in a remaining pool of 77 potential relevant sources.

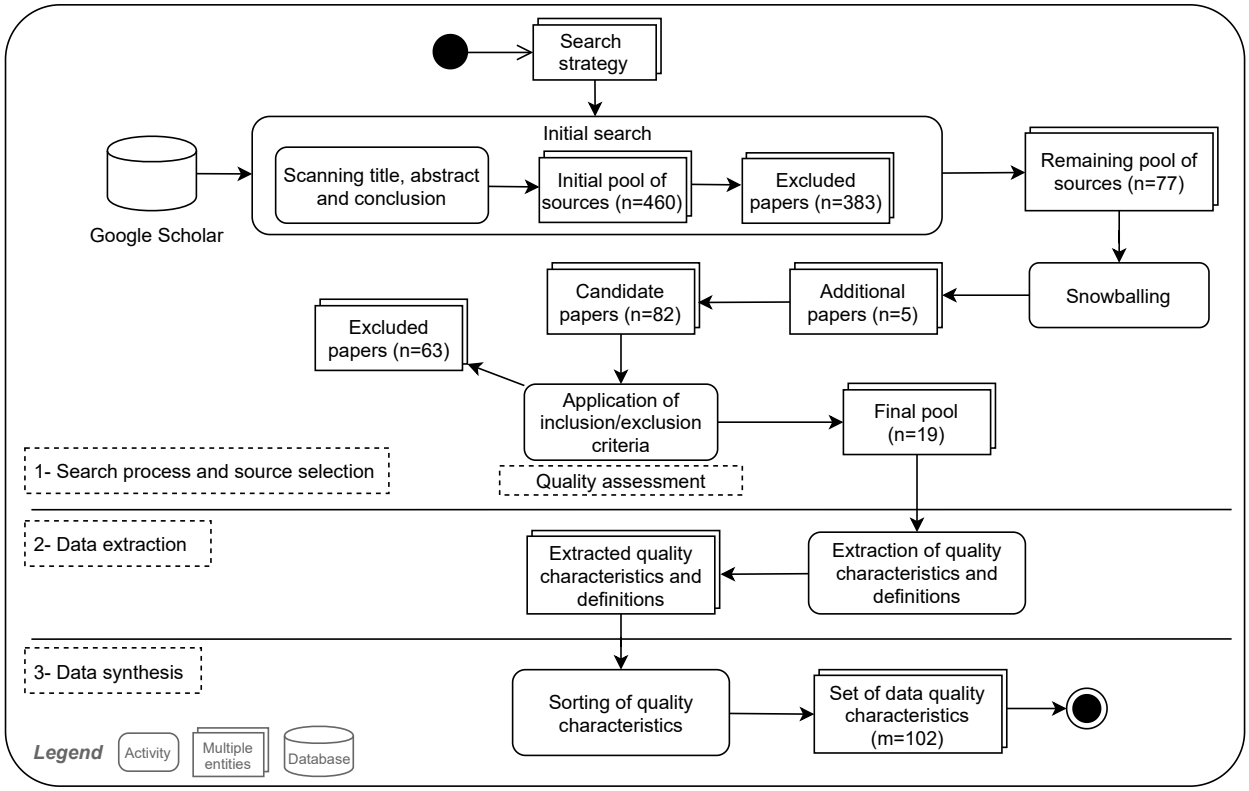

Figure 5: Systematic literature review process.

As a next step we conducted forward and backward snowballing [58] on the 77 identified papers to ensure finding all relevant sources. By examining the reference list of a paper (backward snowballing) and citations to a paper (forward snowballing), we identified 5 additional papers and added them to our set of candidate papers. Thus, we got a final set of 82 candidate papers.

Afterwards, we reviewed each of the 82 papers in detail based on defined inclusion and exclusion criteria shown in Table 1. As a result, we excluded 63 papers and got a final pool of 19 sources for further consideration.

Data Extraction. In the next stage, we extracted relevant quality characteristics from the final pool of papers. Therefore, we used a spreadsheet of Google Sheets[1] and extracted the data store quality characteristics, its definition as well as corresponding grouping or classification schemes (i.e. design & administration quality - schema and data quality - minimality [59]) from the papers. Additionally, we summarized the main content of each quality characteristic's definition in a few key words in a separate column.

Quality characteristics that were explicitly classified as related to inherent data quality (i.e. to data values) in the source papers were not extracted. Moreover, characteristics which were related to the schema of data stores were marked as schema specific. Finally, we got a list of 102 data store quality characteristics (21 of them related to schema) and their definitions. Appendix Appendix B.1 lists the extracted quality characteristics together with the number how often they were extracted. It is worth noting that although characteristics were named identical across sources, they were often defined differently.

Data Synthesis. In the data synthesis stage, we sorted the quality characteristics according to their main stated content. We chose this approach because semantically identical characteristics were often worded

[1] `https://cutt.ly/4jaLZN4`

differently across the sources (e.g. reputation, trustworthiness, verifiability). Through extracting the main content of each characteristic in the data extraction phase and the sorting based on it we ensured that similar characteristics were listed together which kept the further process manageable. In principle, the next step would have been the further synthesis of the characteristics. However, we did not do this at this point in order to be able to assess each of the 102 characteristics individually in the next step with regard to their influence on the inherent data quality.

3.2.2. Relevant Quality Characteristics

As previously stated, not all extracted quality characteristics of data stores have an influence on the quality of their provided data. To identify which characteristics may influence the data quality, we decided to assess their influence on the five most common *inherent data quality characteristics* and on established *data engineering activities*.

Table 2: Data Store Quality Characteristics (sorted alphabetically)

Characteristic	Description
Accessibility	The degree to which data are easily and quickly retrievable.
Availability	The degree to which data are available from a data store.
Completeness	The degree to which a data store is able to represent every meaningful state of the real world.
Contactability	The degree to which a data store provides contact information for further inquiries.
Representational Adequacy	The degree to which a data store presents data in a concise and organized way.
Representational Consistency	The degree to which a data store presents data always in the same format and compatible with previous data.
Security	The degree to which access to data for unauthorized persons is restricted by a data store.
Timeliness	The degree to which a data store provides up-to-date data in a timely manner.
Trustworthiness	The degree to which a data store can be trusted.
Understandability	The degree to which users can understand the data provided by a data store.

For determining the inherent data quality characteristics, we relied on five characteristics provided by the ISO/IEC 25012 data quality model [46] which relate specifically to the extension of data (i.e. *accuracy*, *completeness*, *consistency*, *credibility* and *currentness*). Although sometimes named under different terms (e.g. timeliness instead currentness), these data quality characteristics are widely used in prominent studies on data quality (e.g. [39, 45]) and are therefore well suited to represent the quality of the data provided by data stores.

With regard to the data engineering activities, *data ingestion* (i.e. obtaining and importing data from sources), *data integration* (i.e. combining data from multiple sources), *data storage* (i.e. storing data in repositories), *data cleaning* (i.e. resolving missing data, correcting anomalies) and *data preparation* (i.e. transforming, normalizing or compressing data) were selected as representative data engineering activities. These were selected based on reviewing corresponding literature (e.g. [60, 61, 62]). The idea to assess data store quality characteristics regarding their influence on data engineering activities is based on the following consideration. We assume that data engineers who want to acquire or process data from a data store are more likely to make mistakes when certain quality characteristics of the data store are low. This in turn can lead to a decrease in the quality of the processed data. For example, if metadata or documentation is only available in a limited way, data engineers may introduce errors into the data processing logic which in turn reduces the quality of the processed data.

For the influence assessment, we examined each of the 102 identified data store quality characteristics regarding its potential influence on the inherent data quality or data engineering activities. In detail, we

considered a data store characteristic to be relevant if it either influences at least one of the five inherent data quality characteristics or a data engineering activity. The assessment was done independently by two researchers. In case of different assessments, the corresponding data store quality characteristics were discussed again together and a final assignment was made. An excerpt of the assessment is described in Appendix Appendix B.2.

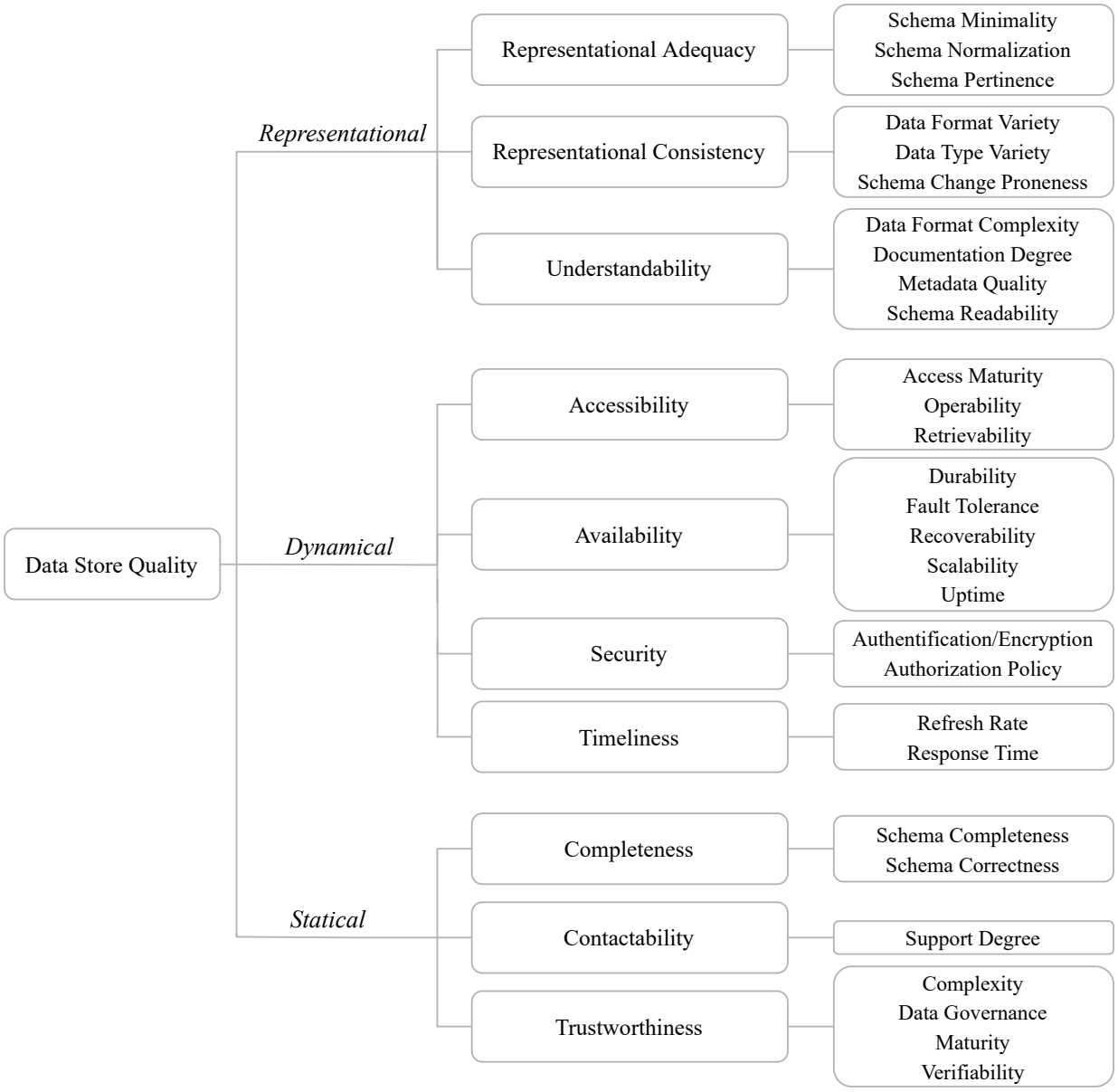

Figure 6: Data store quality model.

In total, we eliminated 17 characteristics resulting in 85 remaining data quality store characteristics (21 of them related to schema) which were assumed to have an influence on the inherent quality of the data provided by a data store. It is worth mentioning that all characteristics that were assumed to influence data engineering activities were also assessed to influence at least one inherent data quality characteristic.

Since the characteristics were already sorted based on their extracted core statement, we now continued by further synthesizing them. In detail, we divided the characteristics into groups based on their core statements. In addition, we assigned each group a quality characteristic name that best represented the main content of the group. An example of the group 'timeliness' is shown in Appendix Appendix B.3. We further derived a definition for each group based on the quality characteristics assigned to it. By doing so, we consolidated the 21 schema-related characteristics into six characteristics. Nevertheless, these characteristics represent only a specific aspect of the quality of a data store (i.e. the schema quality). Therefore, we did not consider them as single quality characteristics, but will take them into account when developing the data store quality model in the next section. Finally, we consolidated the remaining 64 quality characteristics into ten data store quality characteristics which are listed in Table 2.

3.2.3. Quality Model

Our approach to develop the quality model consisted of two steps. First, we grouped the ten identified quality characteristics into superordinate categories. Afterwards, we subdivided the characteristics into further sub-characteristics and properties.

We introduced three parent categories, namely representational, dynamical and statical data store quality. This categorization is partly based on Alvaro et al. [63] who suggest to classify quality characteristics depending whether they are observable at runtime (i.e. *dynamic*) or during the product life-cycle (i.e. *static*). Beside considering the dynamic and static dimensions of the data store quality, we further introduced a category dealing with its *representational* quality dimension. Based on the definition of the ten identified quality characteristics, we assigned each of them to one parent category.

To further subdivide the quality characteristics, we first reviewed the extracted core statements of all characteristics which were assigned to this quality characteristic in Section 3.2.2. Then we used these core statements as properties and sub-characteristics and derived appropriate names and descriptions for each based on literature (e.g. [53, 64, 65]). Additionally, we assigned the six identified schema-related characteristics in Section 3.2.1 as sub-characteristic to a quality characteristic depending on their definition. The final data store quality model is depicted in Figure 6, while the exact descriptions and definitions can be looked up in Appendix Appendix B.4 and Appendix B.5.

3.2.4. Operationalization and Quality Assessment

Based on the work of Punter [66], a checklist was developed to operationalize the developed quality model. The checklist comprises the ten data store quality characteristics and provides a five-point Likert scale (i.e. *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*) for assessing each characteristics. Quality-related statements were defined to represent each characteristic on the checklist. By specifying the agreement to a statement, the level of quality is determined while a high agreement indicates a high level of quality. Through highlighting the main criteria to be considered for each characteristic, the checklist enables a compact assessment. The checklist can be looked up in Appendix Appendix B.6.

To determine the quality of a data store, we assigned points from 0 (i.e. strongly disagree) to 4 (i.e. strongly agree) to each option on the Likert scale. The quality score of a data store ($QS(DataStore)$) can then be computed by summing up all points awarded which can result in a maximum of 40 points ($maxQS(DataStore)$). In the rarely case one quality statement can not be specified, the maximum possible score must be reduced accordingly (i.e. 4 points for each statement not applicable).

We suggest a scale based on Lourenço et al. [67] to interpret the quality scores. The scale subdivides the quality score into five groups (i.e. great, good, average, mediocre, bad), with each group spanning a range of 8 points (i.e. 0-8 *bad*, 9-16 *mediocre*, 17-24 *medium*, 25-32 *good*, 33-40 *great*). This quality score is used to determine the data trustworthiness of data stores in Section 3.4.

3.3. Data Provider Property Catalogue

Figure 7 outlines the different steps of the catalogue development process. In a first step, we conducted an informal literature review to identify *factors* known to influence the data quality in the IIoT. In addition, we gathered *properties* of data providers from the literature that are likely to influence the quality of the

data they provide. Following the review, we synthesized the identified factors and assessed the identified properties regarding their influence on the (inherent) data quality. The remaining properties together with the synthesized factors were then used to develop the data provider property catalogue. Lastly, we developed a checklist to operationalize the catalogue. The following sections (given in Figure 7 in parenthesis) describe each step in more detail.

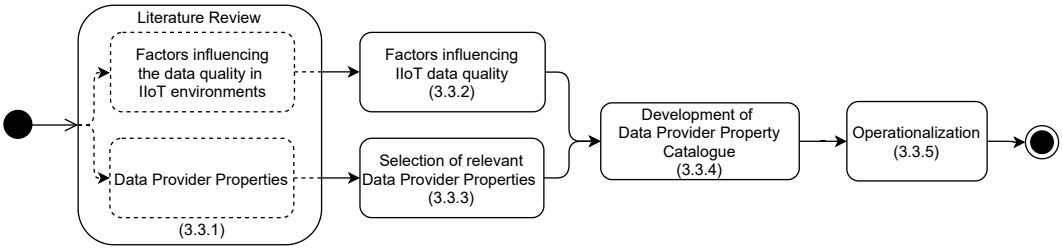

Figure 7: Data provider property catalogue development process.

3.3.1. Literature Review

We conducted the review (corresponding spreadsheets available online[2]) in an informal way for the following reasons. First, a preliminary search revealed that there are already a number of studies reporting on factors that influence the data quality within the IoT. Thus, we decided that it is sufficient to synthesize these contributions. Second, the level of detail of properties of IIoT data providers can be very high. However, such a fine-grained list of properties would only be applicable to a specific type of data provider. Thus, we decided to focus on generic properties that are applicable to a broader range of data providers.

As a first step in the review, we conducted an exploratory search to identify relevant papers. Afterwards, the snowballing technique [58] was applied on this set of papers. As a result, we ended up with 25 candidate papers. In a next step, we reviewed each paper in detail and excluded five papers because their focus was too different (e.g. cloud or big data systems) for our purpose. The remaining 20 sources were then used for the data extraction task.

We extracted both, factors that endanger the IIoT data quality as well as properties of IIoT data providers from eight papers. Six articles each provided only factors or properties. Once all data were extracted, we reviewed all factors as well as properties and eliminated those that were either too abstract (e.g. difficulties of managing data flows) or too detailed (e.g. transmission protocol specific properties). Finally, we got a list of 73 factors and 46 properties.

3.3.2. Factors Influencing IIoT Data Quality

To synthesize and harmonize the 73 identified factors, we relied on the work of Karkouch et al. [23] and Zubair et al. [68]. Based on their classifications, we came up with six main categories compromising 18 factors that influence the IIoT data quality. The six main categories are: *physical environment*, *resource constraints*, *network*, *hardware*, *data processing* and *heterogeneity*. In the following, we briefly describe each category.

Physical Environment. This category refers to the influence of the physical environment on IIoT devices. Extreme environmental conditions (e.g. heat, moisture, dirt) or their dynamic changes (e.g. from very hot to cold) can cause instability of the devices and thus influence the quality of the data provided. Furthermore, damage to the devices due to hostile environments or vandalism can also cause data quality problems.

Network. Networks are an integral part of any IIoT system and can significantly influence the quality of data transmitted. Unreliable network connection (e.g. bandwidth bottlenecks) or latency (e.g. delayed

[2] `https://cutt.ly/6jaZWfT`

transmission) can endanger the currentness of the data provided by IIoT devices. Another common cause of data quality problems are intermittent loss of network connections which can lead to packet loss.

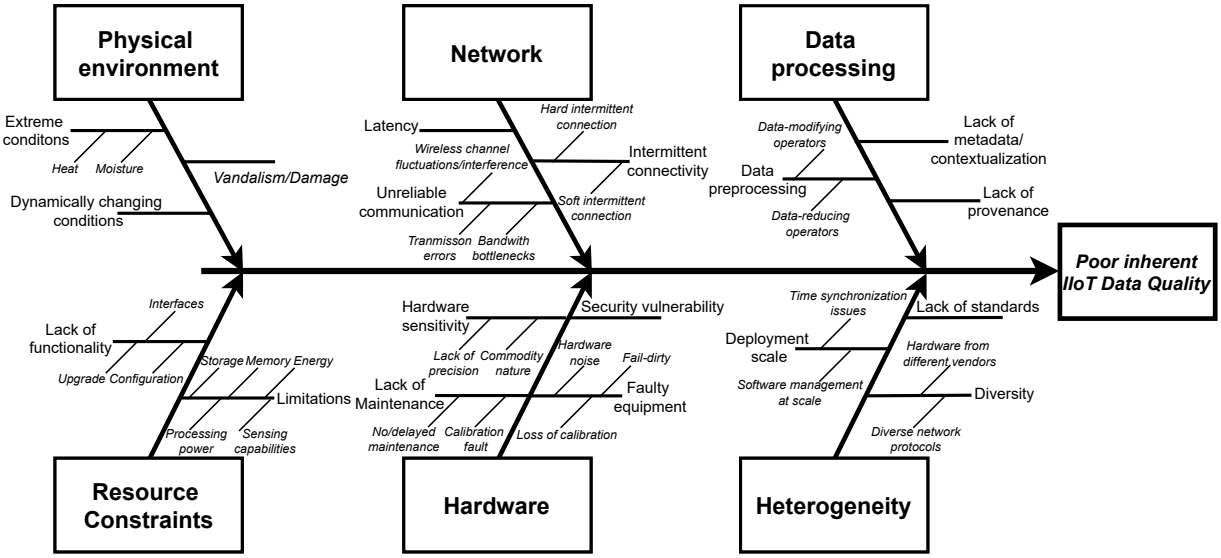

Figure 8: Factors that cause poor inherent IIoT data quality.

Data Processing. Operations on the data or on attributes that describe them (metadata) can have a direct influence on their quality. Since data are commonly processed by IIoT devices to reduce their volume or to preserve privacy, such data-reducing operations (e.g. aggregations, filtering) can influence the data quality through loss of information or metadata. In addition, distributed data processing can cause the inability of deriving the origin of the data (i.e. data provenance) and thus reduce the credibility in the data.

Resource Constraints. IIoT devices often suffer from a lack of resources (e.g. memory, storage, power, computation). Such resource limitations often result in battery outages, bounded processing rates or limited buffers which can affect the quality of the data directly at its origin. In addition, IIoT devices often have limited functionality (e.g. restricted configuration, no updates possible). This lack of functionality may degrade the completeness or accuracy of the data provided.

Hardware. There are several hardware-related factors that definitely have an influence on the quality of the data provided. The commodity nature of many IIoT devices often result in using low-cost hardware. This can result in inaccurate measurements that reduce the accuracy of the data transmitted. Furthermore, such IIoT devices often do not use encryption or other security-relevant protection mechanisms, which makes them vulnerable to attack. This in turn can considerably endanger the quality of the data provided. Moreover, hardware-related failures (e.g. mechanical failures, lost calibration) are often causing massive data quality related issues (e.g. due to erroneous sensor readings). In addition, unmaintained hardware is another potential source of deterioration in the accuracy and credibility of the data provided.

Heterogeneity. The IIoT is characterised by a huge heterogeneity in terms of devices, networks and software stacks. This heterogeneity makes the interaction of devices and software stacks very difficult. Moreover, the lack of standards and the global scale of IIoT implementations often lead to ambiguous contexts and make the enforcement of data quality standards challenging. Therefore, data quality is likely to be negatively influenced due to various problems caused by the heterogeneity of IIoT ecosystems.

If necessary, we renamed the 18 factors in a way that they describe causes of poor data quality. For example, regular maintenance of industrial equipment is a critical factor in ensuring the correctness of the

data they provide. Hence, instead of naming the factor "regular maintenance", we have renamed the factor to "lack of maintenance". This naming enabled us to depict the six main categories as causes and their 18 factors as sub-causes in an Ishikawa (Fishbone) diagram. Figure 8 shows the diagram including further sub-sub-causes (in italics) as concrete examples for most factors.

3.3.3. Relevant Properties

To ensure that only properties are further considered that influence the inherent data quality, we conducted the same assessment procedure as in Section 3.2.2 with one exception. Instead of assessing the properties in terms of their influence on both the inherent data quality and common data engineering activities, we only evaluated them according to the former. This decision was based on the consideration that properties of data providers such as their physical location or kind of power supply are not assumed to have any influence on data engineering activities. However, the assessment resulted in no additional elimination of a property. Thus, all extracted 46 properties were assumed as influencing the quality of the data provided by data providers.

Following the assessment, we consolidated the 46 properties into 12 properties: *age*, *type of data provider*, *environment*, *mobility*, *power source*, *power usage*, *interface mechanism*, *hardware quality*, *maintenance/calibration*, *ease of replacement/repair*, *hardware constraints* and *data preprocessing*.

3.3.4. Catalogue

The catalogue was developed based on the 12 identified properties of data providers and the 18 factors endangering the IIoT data quality. Further, the IIoT analysis framework presented by Boyes et al. [6] was used for structuring the catalogue as it provides categories to characterise IIoT devices.

In detail, we first derived five main categories (i.e. *general*, *location*, *energy*, *connectivity* and *hardware*) based on the categories provided by Boyes et al. and the factors influencing the IIoT data quality. Second, we defined each of the 12 properties in a meaningful way and assigned them to exactly one category. We used the term sensor/device/machine as a representative term for data provider in the definitions. To ensure the coherence of the properties and their corresponding category, we renamed some of the properties accordingly (i.e. connectivity mechanism instead interface mechanism). However, the property *data preprocessing* could not be assigned to a category. Due to its potential influence on the quality of provided data (e.g. compressing data, filtering data or imputing missing data) we decided to add it as the sixth main category. Following, the 12 properties with their definitions and superordinate categories are presented.

- *General:*
  - *Age.* Describes the deterioration of the sensor/device/machine; this includes aspects such as the battery condition, the sensor age and the technology used.
  - *Provider type.* Describes whether the data provided are partly based on human input or were created entirely by the sensor/device/machine itself.
- *Location:*
  - *Environment.* Describes the physical environment in which the sensor/device/machine operates (e.g. temperature range, humidity range, vibration); it also takes into account how often these aspects change (e.g. from extremely hot to cold each day); it further takes into account the level of physical vulnerability (e.g. theft, damage).
  - *Mobility.* Describes whether the sensor/device/machine physically moves during operation.
- *Energy:*
  - *Power source.* Describes the type of power supply of the sensor/device/machine.
  - *Usage.* Describes the power consumption behavior of the sensor/device/machine.
- *Connectivity:*

  - *Mechanism.* Describes the physical mechanism used to convey any data from the sensor/device/ machine.

- *Hardware:*
  - *Quality.* Describes various quality-related hardware characteristics of the sensor/device/ machine; this includes aspects such as the precision, accuracy and sensitivity of the hardware.
  - *Maintenance/calibration.* Describes the frequency of maintenance activities to be carried out on the sensor/device/machine.
  - *Ease of replacement/repair.* Describes how easy it is to repair or replace the sensor/device/ machine in the event of faulty or malfunctioning behavior; this includes aspects such as the physical accessibility and the difficulty of repair, as well as the availability of spare parts.
  - *Constraints.* Describes the extent of resource constraints the sensor/device/machine suffers from; this includes aspects such as power, storage and computation limitations; it also takes into account the availability of backup power supply.

- *Data preprocessing:* Describes the level of data preprocessing that is carried out on the sensor/device /machine; this includes aspects such as data aggregation or data modification.

As a next step in developing the catalogue, we defined *options* for each of the 12 properties. These options provide a means to specify each property. For example, we defined the options *battery* and *hardwired* to specify the *power source* of a data provider. Further, the *connectivity mechanism* of a data provider can be specified by *wireless*, *physical* or *wired*. The final developed catalogue including a detailed description for each option is provided in Appendix Appendix C.1.

3.3.5. Operationalization

To operationalize the catalogue, we developed a checklist similar as for the data store quality model. The checklist comprises all properties of the catalogue together with their options. For conducting the assessment of a data provider, each property has to be specified by one of its options. Through providing a brief description for each option, the checklist enables an intuitive assessment. The checklist can be looked up in the appendix (Appendix C.2).

3.4. Data Trustworthiness Determination

This section describes the determination of the data trustworthiness of data stores, data providers and data sources. The computation is based on the assessments conducted with the developed checklists presented in the Sections 3.2.4 and 3.3.5.

We introduce a trust score τ as a numerical representation of the data trustworthiness. The trust score is defined as a positive number within the range of 0 and 1 and indicates the trust levels as shown in Table 3. The trust score can be computed for data stores, data providers and data sources as outlined in the following.

Table 3: Possible Trust Scores and Their Corresponding Trust Levels

Trust Score τ	Trust Level	Trust Category
1	most trust	high trust
0.9 to 0.99	very high trust	
0.8 to 0.89	high trust	
0.4 to 0.79	moderate trust	medium trust
0.2 to 0.39	low trust	
0.1 to 0.19	very low trust	low trust
0	distrust	

Data Store. The data trustworthiness of a data store is computed based on its calculated quality score (see Section 3.2.4). In detail, the trust score is the quotient of the calculated quality score by the maximum possible quality score as defined in equation 1. This equation reflects our assumption that a data store with high quality (i.e. quality score) has also a high data trustworthiness.

$$\tau(DataStore) = \frac{QS(DataStore)}{\max QS(DataStore)} \tag{1}$$

Data Provider. The data trustworthiness of a data provider is computed based on the data provider property checklist presented in Section 3.3.5. Therefore, points (i.e. 0, 3 or 5) are assigned to every option with which a property can be specified on the checklist. The number 5 is assigned to options that are assumed to represent the highest probability of causing poor inherent data quality. In contrast, the number 0 is assigned to options that indicate the lowest probability of causing poor inherent data quality and thus represent the highest data trustworthiness. Finally, the trust score τ can be computed as one minus the *sum of the points of the selected options* divided by the *maximal score possible* n (i.e. 60) as shown in equation 2. The index i in the equation denotes the number of properties available on the checklist.

$$\tau(DataProvider) = 1 - \frac{\sum selected\_option\_score_i}{n} \tag{2}$$

Data Source. As outlined in Section 3.1, data sources are used as a concept to aggregate several data stores, data providers or any combination of both. To determine the data trustworthiness of a data source, we thus propose to use an averaging trust score $\bar{\tau}$ of the corresponding data stores or data providers. This score represents a holistic assessment of several data stores or several data providers at once, thus with only one checklist. In case that a data source contains both data stores and data providers, we suggest a holistic assessment for each. The trust score τ for a data source can therefore be computed as following:

$$\tau(DataSource) = \frac{\bar{\tau}(DataStore) + \bar{\tau}(DataProvider)}{j} \tag{3}$$

The divisor j in equation 3 is set to 1 when a data source contains only data stores or only data providers. In the case that data stores and data providers are combined, j is set to 2.

3.5. Assessment Procedure

To apply the approach presented, we propose the procedure shown in Figure 9. The procedure is loosely based on the ISO 8000-61 data quality process reference model [69] and consists of five steps, namely *plan*, *model*, *assess*, *determine* and *act*.

Plan. In the first step, the purpose of the assessment is defined. Therefore, consideration should be given to why a data source assessment is being carried out (e.g. identify sources that deliver data of poor quality) or what it is intended to achieve (e.g. reduce the amount of missing data from sensors on the shop floor).

Model. This step deals with modeling the data flow by using the elements (i.e. data store, data provider, data source) and relationships (i.e. delivers/receives data and aggregation) provided by the meta-model. Modeling decisions (e.g. aggregations) should be made based on the defined purpose of the assessment.

Assess. In this step, the modeled elements are assessed based on the defined quality model (data stores, data source), the property catalog (data providers, data source) or both (data source). To conduct the assessments, the presented checklists can be applied.

Determine. This step encompasses the determination of the data trustworthiness for each modeled element by computing a trust score for each. Further, data stores, providers and sources can be assigned to trust levels or categories depending on their calculated trust scores. As the intrinsic properties of data sources can change, a *cyclic repetition* of the 'assess' and 'determine' step can be defined based on the purpose of the assessment.

Act. In the last step, the determined data trustworthiness of each modeled element can be used based on the defined purpose. For example, data providers on the shop floor can be equipped with a fixed power supply to avoid data loss due to frequently empty batteries.

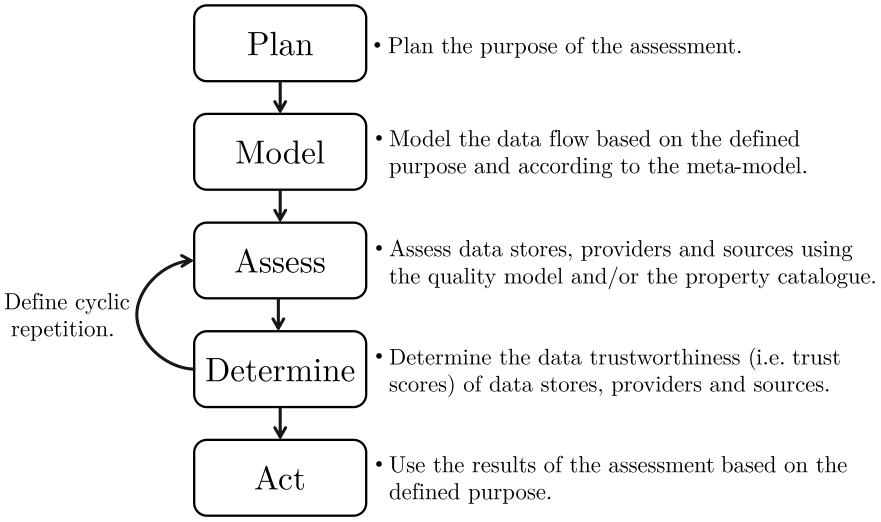

Figure 9: Assessment procedure.

3.6. Application Example

In this section, we theoretically apply the proposed approach to an exemplary industrial use case. The aim of this application example is to illustrate how the proposed approach can be applied.

The presented example is settled within a typical IIoT manufacturing environment consisting of common industrial equipment, machines and sensors. In the concrete use case, a massive amount of data is used to detect product-related failures as early as possible in real-time during production. To achieve this, a machine learning algorithm (e.g. Random Forest) is used to predict the final quality of a product. Depending on the prediction, the products either take different paths through the manufacturing process or are stopped to avoid scrap or rework.

The algorithm is periodically fed with a huge number of data signals coming from the manufacturing environment. Such signals are for example technological parameters provided by machine settings, sensor-measured values of diverse environmental and process-related characteristics or technical parameters of the products measured during in-line quality inspections. In addition, the algorithm uses historical quality information and design parameters of products coming from ERP and MES systems. However, the algorithm suffers from noisy data which leads to a decrease of the prediction accuracy and hence to costly misclassifications. To address this problem, the proposed data source assessment approach is applied based on the previously described assessment procedure (corresponding steps are given in italics in parentheses).

The purpose of the assessment is to identify data sources that are most likely to deliver poor data quality and to evaluate whether the quality of their data can be increased. Alternatively, it should be evaluated whether these data sources can be omitted in the prediction (*plan*).

Figure 10 shows a modeled excerpt of the corresponding manufacturing environment (*model*). For the sake of comprehensibility, only essential components of the data flow are illustrated. As depicted in the figure, databases, the ERP system as well as the MES are modeled as data stores. The machines at the edge tier are modeled either as data providers or data sources.

The next step is to assess the modeled elements, for example with the proposed checklists (*assess*). Afterwards, the data trustworthiness (i.e. trust score) is computed for each modeled data provider, store and source (*determine*). Based on the purpose of use case, a cyclic repetition of the assessment is not needed.

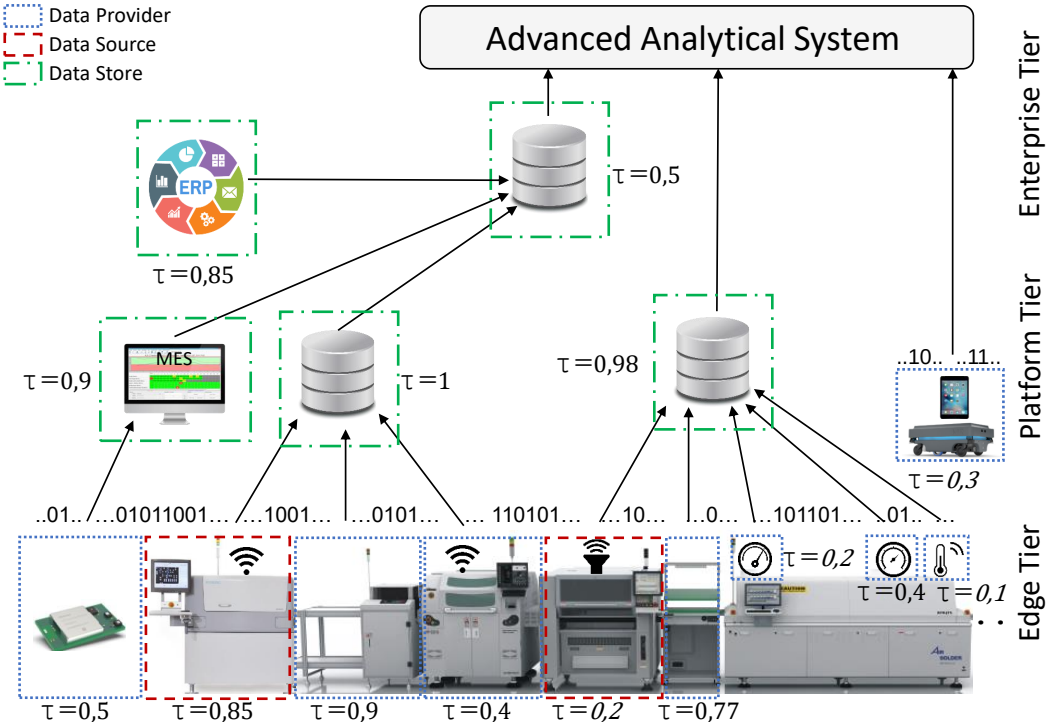

Figure 10: Product quality prediction application example.

The assessment showed (trust score in italics in Figure 10) that two sensors, a machine as well as an automated guided vehicle were assessed as providing data with poor quality. Data engineers can thus examine the data quality from these providers more closely and initiate appropriate measures. For example, data cleaning techniques can be implemented, additional wireless signal repeaters can be installed or the model of the algorithm can be retrained without the data signals of these providers (*act*).

4. Empirical Validation

This section presents an evaluation regarding the validity of the assessments conducted with the presented data source assessment approach. Therefore, we pose the following research question. *Does the approach provide valid assessment results regarding the data trustworthiness of data sources?* To address this research question, we applied a similar research design as Wagner et al. [70] used to evaluate their developed software quality model QuaMoCo. In detail, we first applied the approach to industrial data sources and conducted an assessment with the developed checklists. Afterwards, experts were asked to assess the quality of the data provided by these data sources. Finally, the results were compared and analyzed regarding their consistency. Following, we describe the setting of the validation, its execution and then the results.

4.1. Setting

We applied the developed approach to ten randomly selected data sources of an Austrian electronic manufacturing services company as listed in Table 4. The selected data sources are directly related to the company's production environment where complex electronic modules, components and systems are manufactured. In detail, we selected the following five machines from three Surface Mount Technology (SMT) manufacturing lines: a solder stencil printer, two fully automated SMT assembly machines of different generations, an automatic optical inspection machine and a solder reflow oven. SMT manufacturing describes a process where electronic components are assembled on printed circuit boards. In addition to the five machines, we selected three IT systems as data sources: an IT system for quality and inspection data

acquisition, an ERP system and a software for managing SMT component shapes. The final two selected data sources were a wearable barcode scanner combined with a smartphone and a nitrogen sensor. The left side of Figure 11 shows a part of the company's manufacturing lines with a solder reflow oven, an assembly machine and a solder stencil printer highlighted with arrows. On the right side of the figure, three data sources used for the assessment are shown.

Table 4: Overview of Data Sources

#	Description
1	Solder stencil printer
2	SMT assembly machine
3	SMT assembly machine
4	Automatic optical inspection machine
5	Solder reflow oven
6	ERP system
7	Quality & inspection data system
8	SMT component shape library software
9	Wearable barcode scanner & smartphone
10	Nitrogen sensor

Table 5: Overview of Experts

#	Role
1	IT technician
2	IT system administrator
3	Software/data engineer
4	Technical production leader
5	Process engineer

4.2. Execution

To assess the ten selected data sources, they were first modeled as data stores, data providers or data sources according to the developed approach. Based on the modeling, each real-world data source was assessed with the corresponding checklist(s). As an illustration, the data provider assessment of the wearable barcode scanner combined with a smartphone can be looked up in Appendix Appendix D.1.

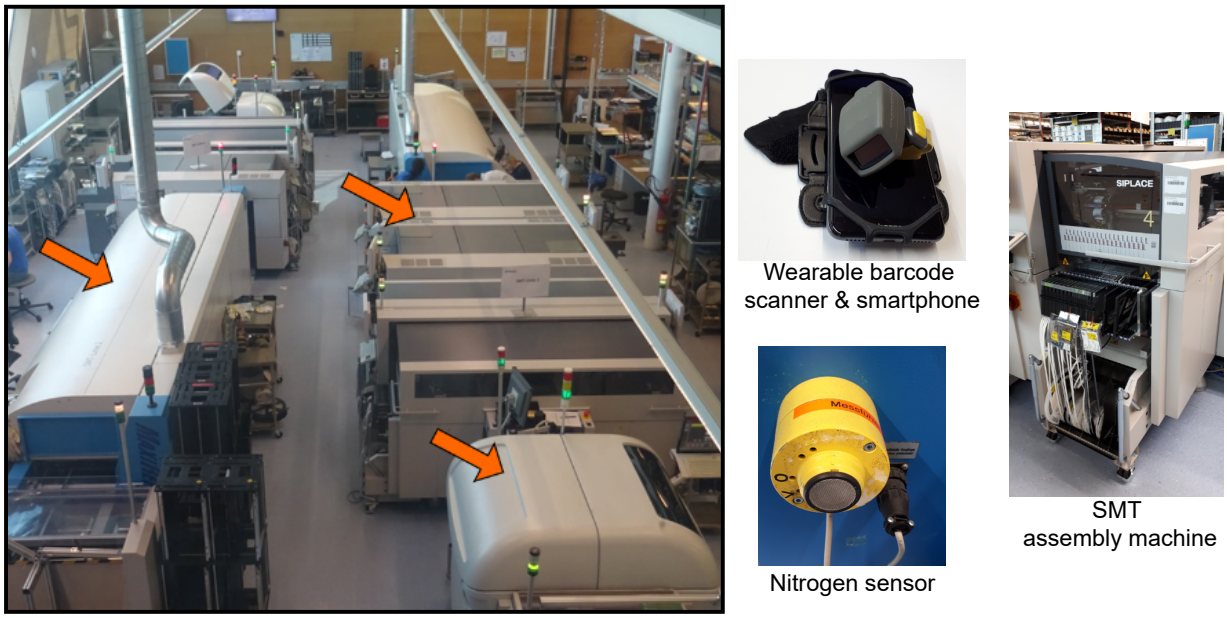

Figure 11: Picture of a part of the company's manufacturing lines and three assessed data sources.

Five experts with at least five years of experience in the company were then asked to assess the quality of the data provided by each data source. The experts were selected based on their background and practical experience regarding the data sources as shown in Table 5. In fact, we selected three experts based on their

IT-related experience with the data sources (e.g. data engineering activities). The two remaining experts were selected because they frequently use the data from the data sources in their daily work and are therefore well informed about their quality.

The assessment of the experts was then carried out using Google Forms. Each expert had to assess the data quality of each data source based on an ordinal scale (i.e. excellent data quality, good data quality, low data quality, poor data quality and can not state).

4.3. Results

Before we started analysing the results, we first measured the agreement among the experts regarding their assessments. Therefore, we calculated Krippendorff's alpha α [71, 72]. We used this coefficient instead of others (e.g. Cohens's kappa, Fleiss' kappa) for the following reasons. First, there were missing quality assessments (i.e. 'can not state' ratings). Second, there were multiple raters (i.e. five experts) and third, their composition remained the same during the whole assessment. The latter is an often overlooked requirement of Fleiss' kappa. We calculated a value of 0.79 for α. Krippendorff [73] suggest $\alpha \geq 0.667$ for tentative conclusions and $\alpha \geq 0.8$ as an acceptable level of agreement. Thus, we assumed that the experts' assessments were consistent. A visualization of the consistency of the assessments by the experts is shown in Figure 13.

In total, 44 valid quality assessments were given whereas six times experts chose the option 'can not state'. A detailed overview about the assessments per expert as well as per data source can be looked up in the appendix (Appendix D.2 and Appendix D.3). Because the experts used an ordinal scale for their assessment, we decided to analyse the consistency between the assessments of the experts and the developed approach based on the absolute ranking provided by each. Therefore, we calculated the Spearman's rho correlation coefficient which gave a value of 0.69 with a p-value of 0.04. According to Taylor [74], these values indicate a significant moderate correlation between the ranking of the data sources provided by the experts and the ranking of the data sources provided by the developed approach. Descriptive statistics of the expert assessments and details of the conducted calculations are provided online[3]. The trust scores calculated with the data source assessment approach are depicted in Appendix Appendix D.4.

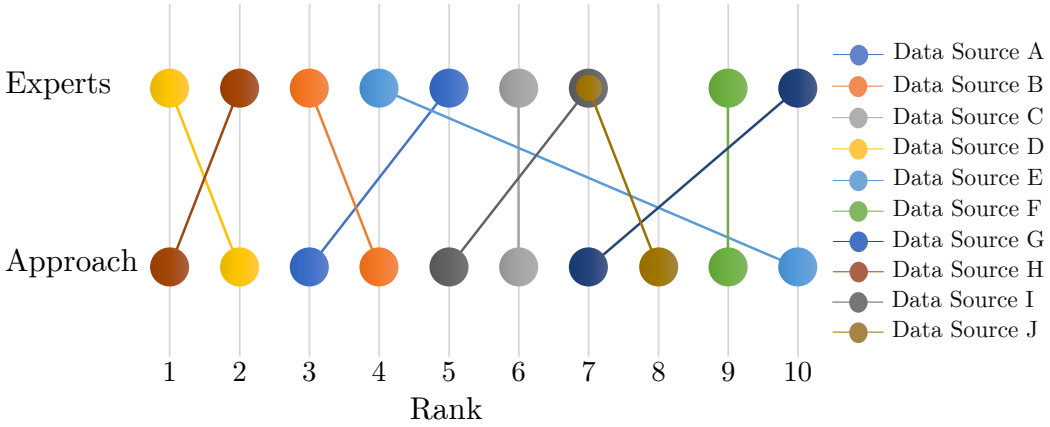

Figure 12: Comparison of data source rankings based on expert assessments and the developed approach.

Figure 12 illustrates the rankings of the ten data sources based on both assessments. Due to confidentiality reasons, brand names and descriptions were omitted and the data sources were numbered alphabetically (i.e. Data Source A, Data Source B and so forth). The majority of the data sources were either ranked equally by the approach or ranked one rank different compared to the ranking provided by the experts. Interestingly, one data source (Data Source E) was rated by the approach with the lowest data trustworthiness, whereas the experts rated it as one of the half of the data sources providing better data quality

[3] `https://cutt.ly/jjaK40Y`

(rank 4). This discrepancy could be partially attributed to the fact that the computed trust score τ of the approach for data source E is close to the τ of the data sources ranked nearby. Further, it has to be attributed that data source E was the data source with the highest disagreement among the experts (see Figure 13 and Appendix Appendix D.3).

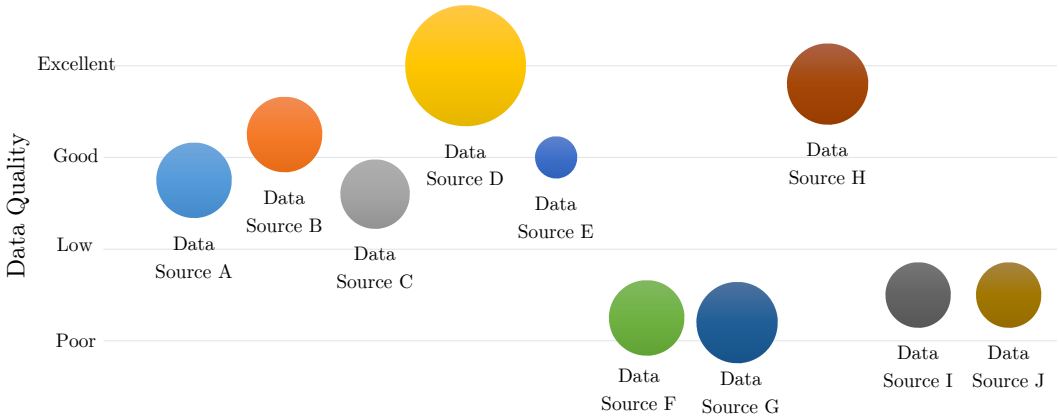

Figure 13: Consistency of the experts on the quality of the data provided by each data source. The larger the bubble diameter, the more the experts agreed on the data quality of a source (the bubble of Data Source D represents a full agreement). The depicted data quality of a source (y-axis) is the average data quality resulting from all expert assessments of a source.

Taken together, the found moderate correlation suggest that the assessments conducted with the data source assessment approach can be treated as valid.

5. Discussion

The main objective of this article was to develop an approach to determine the probability that IIoT data sources provide correct data (i.e. data trustworthiness). In the case study, it was found that assessments conducted using the developed approach correlated moderately with assessments conducted by experts. Following, Section 5.1 briefly discusses the approach in connection with the previously mentioned related work (see Section 2.3). Afterwards, we interpret the result of the case study (Section 5.2) and then discuss threats to validity in Section 5.3.

5.1. Related Work

The developed data source assessment approach is based on the idea to use assessments of data sources to infer their data trustworthiness. Therefore, the approach puts its emphasis on the intrinsic characteristics of data sources to reason about the quality of their data provided. This is in contrast to previous work on data trustworthiness (e.g. [17, 15, 51]), which mainly assess the trustworthiness on the basis of data values. Since data values typically change dynamically, these approaches are usually based on an ongoing cyclical determination of the trustworthiness of the data. As the intrinsic characteristics of data sources do not change as dynamically as data values, our approach does not require an ongoing trustworthiness determination. However, repetition cycles (i.e. reassessments of data sources) can be defined depending on the specific application scenario.

In addition, the presented approach differs from work related to data source assessment (e.g. [52, 54]) in following ways. First, it provides a concise way to distinguish between data stores and data providers. Second, to the best of our knowledge, it is the first approach that focuses on IIoT data sources.

5.2. Interpretation

The finding of the case study confirms that the intrinsic characteristics and properties of data sources may influence their provided data quality. This is in line with previous work, which describes the influence of various factors on data quality in the IoT context (e.g. [23, 68]). Additionally, the results are consistent with literature claiming that characteristics associated with the quality of data stores (e.g. [22, 75]) are able to influence the quality of the data stored.

Additionally, the case study revealed that the meta-model developed is practicable in modeling data sources within real-world manufacturing environments. A note of caution is due here since it is difficult to provide specific criteria when to model a real-world data source as data store or data provider. The provided abstraction of a data source in the meta-model, containing characteristics of data stores as well as properties of data providers, therefore enables the combined use of both concepts. Nonetheless, as the focus of the case study was to validate the assessments results and not the meta-model, we cannot guarantee that the data source abstraction element and its data trustworthiness calculation is appropriate for large-scale data flows.

With regard to the generalization of the meta-model, it can be assumed that its key concepts (i.e. data provider, data store) and relationships (i.e. delivers/receives data) are domain-agnostic. In fact, the concept of a data store and the proposed quality model can be applied to a wide range of domains. As typical entities in information technology, data stores have the same properties regardless of the domain used. However, although the concept of a data provider can be generalized to other domains, the presented property catalogue of data providers is limited in terms of its application to the industrial domain. This is because the properties of data providers are domain-dependent. For example, a sensor has different properties than a statistical agency acting as a data provider.

5.3. Threats to Validity

To avoid common threats to validity, we developed the approach in a documented and systematic way based on existing literature. We further provide all intermediate results of the development (e.g. details on the literature review, data quality characteristic assessments) online to ensure transparency and traceability. The case study presented in this article was conducted to empirically validate the approach. However, according to Runeson and Höst [76], case study research involves some threats to validity. Following, we describe threats to construct validity, internal validity, reliability, external validity [76] as well as conclusion validity [77] and our applied countermeasures.

To minimize threats to *construct validity*, the assessment forms used by the experts provided a definition of data quality based on five inherent data quality characteristics. The experts were instructed to contact us if anything was unclear or if they had any questions during the assessment. Further, the checklists used to apply the approach contained definitions of all data store quality characteristics as well as all data provider properties. *Internal validity* was addressed by calculating Krippendorff's alpha. The calculated value of alpha suggests that the expert-based assessments of the data sources are consistent. Thus, we safely assume that the expert assessments adequately represent the data trustworthiness of these sources. Another threat to internal validity is that some data store characteristics and data provider properties that have a significant influence on data quality may not be included in the approach. To mitigate this threat, we applied the snowballing technique in addition to the keyword search during the literature review. To ensure *reliability* of the case study, the forms used for the assessment, the checklists as well as all data processing tasks were well documented and are provided online. In terms of *external validity*, the results need to be interpreted with caution because the empirical validation was limited in terms of the number of data sources and the specific case context. Following, we cannot conclude that the developed approach would also provide valid results in other domains (e.g. finance) or larger industrial scales. Finally, to inhibit threats to *conclusion validity* we calculated the p-value of the correlation coefficient and set the significance level to 5 % to minimize the probability that the results occurred by chance. A further threat to conclusion validity is the usage of an ordinal ranking scheme in the case study. Thus, conclusions drawn based on the absolute distances between the trust scores of data sources should be treated with caution.

6. Conclusions

In this article, we presented an approach to assess IIoT data sources to infer the trustworthiness of their provided data. First, we introduced a meta-model which enables a decomposition of data sources into data stores and data providers. Subsequently, a quality model was proposed that comprises characteristics related to the intrinsic quality of data stores (e.g. availability, metadata quality) to determine their data trustworthiness. Further, a catalogue based on properties of data providers was presented to infer the trustworthiness of their provided data. To operationalize the quality model and the catalogue, checklists were proposed. We then presented the concrete calculation of the data trustworthiness for data sources based on these checklists. A conducted industrial case revealed that developed approach is able to provide a valid ranking of data sources regarding their data trustworthiness compared to an expert-based assessment (Spearman's rho correlation coefficient of 0.69 with a p-value of 0.04).

Overall, this article strengthens the idea that the intrinsic characteristics of data sources play a significant role in determining the trustworthiness of the data they provide. The presented approach contributes to the field of data trustworthiness and data source assessment in following ways. First, the proposed characteristics and properties to reason about the data trustworthiness of data sources are valuable for other approaches in the field of data trustworthiness. In fact, existing data trustworthiness techniques can apply the proposed characteristics and properties to take data sources into more account within their computations. Second, the proposed meta-model is a suitable means for developing data source assessment approaches in other areas. Further, other researchers are able to apply the data store quality model within their area of research based on its domain-independence.

This article further contributes to existing knowledge of data trustworthiness in IIoT environments by several ways. First, the comprehensive overview about factors influencing the data quality within the IIoT will be of interest by researchers investigating industrial data quality problems. Second, the proposed data provider property catalogue is a valuable and easy to use tool for practitioners to assess industrial data sources.

The presented approach will further prove useful in steering data validation activities (e.g. [56]). Thereby, data validation can be focused on data coming from data sources with low data trustworthiness. Nevertheless, the approach also can be used for data quality improvement programs or for risk management initiatives.

We intent to focus our future research on the development of a tool support which enables the modeling and assessment of data sources in a more comprehensible way. Further, we aim to target further work on validating the meta-model as well as to apply the approach in another, larger industrial setting. In addition, further research should be carried out to investigate the extent to which each data source characteristic influences the data quality. On this basis, weights can be assigned to all characteristics within the approach, which would improve the validity of the assessment results. Future work is further needed to investigate the applicability of the meta-model to large-scale data flows. Especially, the data trustworthiness calculation for data sources receiving data from a lot of data providers should be examined.

Acknowledgement

The research reported in this article has been partly funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET - Competence Centers for Excellent Technologies Programme managed by Austrian Research Promotion Agency FFG.

Appendix A. ISO/IEC 25012 Data Quality Model

Table A.6: ISO/IEC 25012 Data Quality Model [46]

Characteristic	Description	Categ.[4]
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.	Inherent
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.	Inherent
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use.	Inherent
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.	Inherent
Currentness	The degree to which data has attributes that are of the right age in a specific context of use.	Inherent
Accessibility	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.	In.&Sys.
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.	In.&Sys.
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.	In.&Sys.
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.	In.&Sys.
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.	In.&Sys.
Traceability	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.	In.&Sys.
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.	In.&Sys.
Availability	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.	System
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.	System
Recoverability	The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.	System.

[4]Data Quality Category (Categ.): System-Dependent (System), Inherent and System-Dependent (In.&Sys.)

Appendix B. Data Store Quality Model

Appendix B.1. Extracted Quality Characteristics

Table B.7: Extracted Quality Characteristics - Schema-related Characteristics shown in Italics

Characteristic	#	Characteristic	#	Characteristic	#
Accessibility	8	Data Governance	1	Timeless	1
Security	5	Data Specifications	1	Timeliness & Punctuality	1
Reliability	4	Durability	1	Traceability	1
Availability	3	Ease of Operation	1	Transactional Availability	1
Completeness	3	Efficiency	1	Trustworthiness	1
Reputation	3	Freshness	1	Trustworthiness & Verifiability	1
Value-Added	3	Interactivity	1	Understandability	1
Customer Support	2	License	1	Usage	1
Documentation	2	Metadata	1	Value of Tail	1
Ease of Manipulation	2	Openness	1	Verifiability	1
Interpretability	2	Organization	1		
Relevance	2	Origin	1	*Minimality*	4
Access Security	1	Performance	1	*Completeness*	4
Accuracy & Reliability	1	Portability	1	*Correctness*	3
Amount of Data	1	Presentation	1	*Readability*	2
Attractiveness	1	Refresh Rate	1	*Consistency*	1
Authority & Sustainability	1	Representational Adequacy	1	*Relevance*	1
Clarity	1	Representational Consistency	1	*Interpretability*	1
Concise Representation	1	Response Time	1	*Level of Detail*	1
Confidentiality	1	Responsiveness	1	*Scope*	1
Consistent Representation	1	Retrievability	1	*Ability of Integration*	1
Contactability	1	Significance	1	*Normalization*	1
Credibility	1	System Availability	1	*Pertinence*	1

Appendix B.2. Data Characteristics Assessment Procedure

Table B.8: Data Characteristics Assessment Procedure (Excerpt)

Source	Characteristic	Definition	Inherent DQ					Data Eng.
			Accur.	Comp.	Consi.	Credi.	Curre.	
Caro et al.	Organization	The organization, visual settings or typographical features (colour, text, font, etc.) and the consistent combinations of these various components.	-	-	-	-	-	-
Wang & Strong	Accessibility	The extent to which data are available or easily and quickly retrievable.	-	x	-	-	x	x
Jarke et al.	Security	Describes the authorization policy and the privileges each user has for the querying of the data.	-	-	-	x	-	-

Visual settings or typographical features of data stores cannot influence the quality of the stored data. Hence, the characteristic *organization* was assessed as non-influencing characteristic. *Accessibility* was assessed as influencing the (inherent) data quality characteristics completeness and currency based on following considerations. Non-available or delayed data cannot be of the right age and hence reduces the *currentness* inherent data quality characteristic. Further, if data are not available it is apparent that the actual data values are missing and therefore reduces the completeness of the data. Moreover, due to its definition, low accessibility is associated with complex and difficult retrievability which in turn can lead to an increase in data ingestion or integration mistakes. In case a data store provides authorization policies (i.e. *security*), it increases the credibility of the data provided. However, data engineering activities are not influenced because there is no increased likelihood of making data-related mistakes caused by authorization policies.

Appendix B.3. Data Quality Characteristics Consolidation

Table B.9: Consolidation of Timeliness Characteristic

Source	Characteristic	Definition
Caro et al.	Response Time	Amount of time until complete response reaches the user.
Stróżyna et al.	Timeliness & Punctuality	Data update (time interval between the occurrence of an event and availability of the data which describe it) and time delay in publishing updated information.
Jarke et al.	Responsiveness	Is concerned with the interaction of a process with the user (e.g. a query tool which is self reporting on the time a query might take to be answered).
Assaf, Senart	Performance	Is the data source capable of coping with increasing requests in low latency response time and high throughput?
Rogova, Bosse	Credibility	Is the frequency, with which a process and model, or a human agent produce a correct answer.
Zhu, Buchmann	Refresh Rate	Refers to the timeliness with which data is posted to the site, but it also means that volatile data that is overwritten at a fast refresh rate must be extracted at the same rate to avoid loosing data.
Rafique et al.	Efficiency	The degree to which information has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.
Firmani et al.	Freshness	The freshness of a source at a time t is the probability that a randomly selected entity is up-to-date.
Rogova, Bosse	Timeless	The information is presented by the time it must be used and whether the dynamics of information in the real world is reflected by the dynamic of information presentation (when information is presented).

Appendix B.4. Detailed Quality Model

Data Quality Category	Data Quality Characteristic	Description	Properties and Data Quality Sub-characteristics
Representational Data Store Quality It emphasized the importance of the representational role of system; that is, the system must present data in such a way that they are interpretable, easy to understand, and concisely and consistently represented.	***Representational Adequacy***	Degree to which a data store presents data in a *concise and organized way*.	**Schema Minimality** **Schema Normalization** **Schema Pertinence**
	Representational Consistency	Degree to which a data store presents data *always in the same format and compatible with previous data*.	**Data Format Variety** **Data Type Variety** **Schema Change Proneness**
	Understandability	Degree to which *users can understand the data* provided by a data store.	**Data Format Complexity** **Documentation Degree** **Metadata Quality** **Schema Readability**
Dynamical Data Store Quality It emphasized the importance of the dynamic role of system; that is, the system must be secure, available, accessable and provide up-to-date data with high performance.v	***Accessibility***	Degree to which data are *easily and quickly retrievable*.	**Access Maturity** **Operability** **Retrievability**
	Availability	Degree to *which data are available* from a data store.	**Durability** **Fault Tolerance** **Recoverability** **Scalability** **Uptime**
	Security	Degree to which *access to data for unauthorized persons is restricted* by a data store.	**Authentification/Encryption** **Authorization Policy**
	Timeliness	Degree to which a data store *provides up-to-date data in a timely manner*.	**Refresh Rate** **Response Time**
Statical Data Store Quality It emphasized the importance of the static role of system; that is, the system must be complete and trustworth as well as provide support capabilities.	***Completeness***	Degree to which a data store is able to represent *every meaningful state of the real world*.	**Schema Completeness** **Schema Correctness**
	Contactability	Degree to which a data store provides *contact information for further inquiries*.	**Support Degree**
	Trustworthiness	Degree to which a *data store can be trusted*.	**Complexity** **Data Governance** **Maturity** **Verifiability**

Figure B.14: Data store quality model - detailed.

Properties and Data Quality Sub-characteristics	Description
Schema Minimality	Degree to which the schema is modeled without redundancies.
Schema Normalization	Degree to which the schema is de-/normalized.
Schema Pertinence	Degree to which the schema is modeled without unnecessary elements.
Data Format Variety	Refers to the amount of different data formats used by a data store to present the data.
Data Type Variety	Refers to the amount of different data types used by a data store to present the data.
Schema Change Proneness	Degree to which the schema is likely to change in the future.
Data Format Complexity	Describes the overall complexity of the data format used by a data store to represent the data; it includes aspects such as the amount of proprietary/open data formats used and the possibility to process the data format directly (i.e. no transformations needed); it also takes into account whether the data are presented in a clear, simple and well structured way.
Documentation Degree	Degree to which a data store provides useful and complete documentation.
Metadata Quality	Degree to which a data store provides metadata that are available, complete, useful, interpretable and human readable; it also takes into account the availability of data dictionaries or repositories.
Schema Readability	Degree to which the schema is modeled in a natural, clear and self-explanatory way.
Access Maturity	Degree to which provided access methods (e.g. APIs) and protocols of a data store are mature and reliable.
Operability	Degree to which a data store provides possibilities (e.g. management system, navigation mechanisms) that guides the access to it.
Retrievability	Degree to which query possibilities of a data store are mature, reliable and require no additional requirements (e.g. registration).
Durability	Degree to which a data store keeps the data available for retrieval.
Fault Tolerance	Degree to which a data store operates as intended despite the presence of hardware or software faults (e.g. CAP, ACID, BASE).
Recoverability	Degree to which a data store can re-establish the desired operating state after an interruption or failure (e.g. CAP, ACID, BASE).
Scalability	Degree to which a data store can deal with varying workloads without downtimes.
Uptime	Degree to which a data store is up and running, i.e. reachable (e.g. CAP, ACID, BASE).
Authentification/Encryption	Degree to which authentification and encryption are ensured by a data store.
Authorization Policy	Degree to which authorization policies are in place on a data store.
Refresh Rate	Degree to which the time interval between the occurence of an event and the availability of the data (which describe this event) on a data store sufficient.
Response Time	Degree to which the response and processing times of a data store are sufficient.
Schema Completeness	Degree to which all real-world concepts are represented in the schema (complete representation).
Schema Correctness	Degree to which the schema corresponds to the real-world it is supposed to model (correct representation).
Support Degree	Describes the overall degree of support; it includes aspects such as service agreements, availability of previous developers and domain experts.
Complexity	Describs the overall complexity of a data store; it includes aspects such as the complexity of the data model, the schema size and the extent of data processing/transforming; it also takes into acount whether the data store type is known to the current developers.
Data Governance	Degree to which data governance principles (e.g. data policies, cleaning, backup etc.) are applied on a data store.
Maturity	Degree to which a data store is reliable (i.e. not prone to failures), stable (i.e. no recent upgrades, changes) and current (i.e. latest stable software release installed).
Verifiability	Degree to which a data store provides information on the data origin, on changes made to the data and possibilities to check the data for correchness.

Figure B.15: Data store quality model - sub-characteristics.

Appendix B.6. Data Store Quality Checklist

Data Store Quality Characteristic	Rating				
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Representational Data Store Quality					
Representational Adequacy	☐	☐	☐	☐	☐
The data store presents data in a concise and organized way.					
Criteria (e.g.): *schema has no redundancies or unnecessary elements and is sufficiently de-/normalized*					
Representational Consistency	☐	☐	☐	☐	☐
The data store presents data always in the same format and compatible with previous data.					
Criteria (e.g.): *few different data formats and types used, schema is not likely to change*					
Understandability	☐	☐	☐	☐	☐
The users can easily understand the data provided by the data store.					
Criteria (e.g.): *useful documentation and metadata available, schema is natural and clear, simple, open and well structured data formats used*					
Dynamical Data Store Quality					
Accessibility	☐	☐	☐	☐	☐
The data can be retrieved easily and quickly from the data store.					
Criteria (e.g.): *access, query methods and protocols are mature, guided access (e.g. management system)*					
Availability	☐	☐	☐	☐	☐
The data store is operational and accessible all the time.					
Criteria (e.g.): *high fault tolerance, recoverability, scalability and uptime, data are permanent retrievable*					
Security	☐	☐	☐	☐	☐
The data store restricts access to data for unauthorized persons.					
Criteria (e.g.): *authentification, encryption and authorization used*					
Timeliness	☐	☐	☐	☐	☐
The data store provides up-to-date data in a timely manner.					
Criteria (e.g.): *high refresh rate, response and processing times*					
Statical Data Store Quality					
Completeness	☐	☐	☐	☐	☐
The data store represents every meaningful state of the real world.					
Criteria (e.g.): *schema is complete and correct modeled*					
Contactability	☐	☐	☐	☐	☐
The data store provides contact information for further inquiries.					
Criteria (e.g.): *service agreements and domain experts available*					
Trustworthiness	☐	☐	☐	☐	☐
The data store is a reputable and credible source of data.					
Criteria (e.g.): *data policies, backups, pertinance in place, simple data model, familiar data store type, data store is reliable, stable and current*					

Figure B.16: Data store quality checklist.

Appendix C. Data Provider Property Catalogue

Appendix C.1. Detailed Data Provider Property Catalogue

Category	Property	Description	Option	Additional Description	Points
General	Age	Describes the deterioration of the sensor/device/machine; this includes aspects such as the battery condition, the sensor age and the technology used.	**old**	legacy technology, operating hours exceed average lifespan	5
			middle	state-of-the art technology, operating hours lower average lifespan	3
			new	new technology, few operating hours	0
	Provider type	Describes whether the data provided are partly based on human input or were created entirely by the sensor/device/machine itself.	**persons involved**	(partial) manual data entry/dependence	5
			only sensor/device/machine	no manual data entry/dependence	0
Location	Environment	Describes the physical environment in which the sensor/device/machine operates (e.g. temperature range, humidity range, vibration); it also takes into account how often these aspects change (e.g. from extremely hot to cold each day); it further takes into account the level of physical vulnerability (e.g. theft, damage).	**harsh, dynamically changing**	extreme conditions, conditions change often, likely to damage/theft	5
			moderate, seldom changing	moderate conditions, conditions change seldom, damage/theft is possible	3
			mild, never changing	mild conditions, conditions change never, damage/theft is not likely	0
	Mobility	Describes whether the sensor/device/machine physically moves during operation.	**always**	moving most of the time	5
			sometimes	moving sometimes	3
			never	never moving	0
Energy	Power source	Describes the type of power supply of the sensor/device/machine.	**battery**	battery-powered	5
			hardwired	wired power supply	0
	Usage	Describes the power consumption behavior of the sensor/device/machine.	**low**	rarely on, passive mode/sleep mechanisms used	5
			medium	power saving, passive mode/sleep mechanisms used	3
			high	always on, no passive mode/sleep mechanisms used	0
Connectivity	Mechanism	Describes the physical mechanism used to convey any data from the sensor/device/machine.	**wireless**	spectrum, light, sound; e.g. radio frequency (e.g. RFID, Bluetooth Low Energy)	5
			physical	pressure, mechanical, thermal; e.g. ultrasonic communication (e.g. SoniTalk)	3
			wired	electrical, fibre optic; e.g. Ethernet	0
Hardware	Quality	Describes various quality-related hardware characteristics of the sensor/device/machine; this includes aspects such as the precision, accuracy and sensitivity of the hardware.	**low**	low cost sensors, commodity nature, not suitable for use in the application used	5
			medium	state-of-the art quality, suitable for use in the application used	3
			high	highly precise and robust hardware	0
	Maintenance/ calibration	Describes the frequency of maintenance activities to be carried out on the sensor/device/machine.	**never**	never serviced	5
			sporadic	sporadic serviced (e.g. in case of failures)	3
			regulary	regulary serviced	0
	Ease of replacement/ repair	Describes how easy it is to repair or replace the sensor/device/machine in the event of faulty or malfunctioning behavior; this includes aspects such as the physical accessibility and the difficulty of repair, as well as the availability of spare parts.	**hard**	unsupported, no experience, no access	5
			medium	supported, no experience, accessable	3
			easy	supported, knowledge available, accessable	0
	Constraints	Describes the extent of resource constraints the sensor/device/machine suffers from; this includes aspects such as power, storage and computation limitations; it also takes into account the availability of backup power supply.	**severe limitations**	low battery capacity, low-performance hardware, no backup power source	5
			moderate limitations	state-of-the art hardware	3
			no limitations	high battery capacity, high-performance hardware, backup power available	0
Data preprocessing		Describes the level of data preprocessing that is carried out on the sensor/device/machine; this includes aspects such as data aggregation or data modification.	**high**	a lot of data preprocessing carried out	5
			medium	a state-of-the art level of preprocessing carried out	3
			low/no	no essential data preprocessing carried out	0

Figure C.17: Data provider property catalogue- detailed.

Appendix C.2. Data Provider Property Checklist

Properties	Choice		
General			
Age Choose the age of the sensor/device/machine.	☐ **old** legacy technology, operating hours exceed average lifespan	☐ **middle** state-of-the art technology, operating hours lower average lifespan	☐ **new** new technology, few operating hours
Provider type Choose whether the provided data are partially based on human input or completely created by the sensor/device/machine itself.	☐ **persons involved** (partial) manual data entry/dependence		☐ **only sensor/device/machine** no manual data entry/dependence
Location			
Environment Choose the physical environment in which the sensor/device/machine operates.	☐ **harsh, dynamically changing** extreme conditions, conditions change often, likely to damage/theft	☐ **moderate, seldom changing** moderate conditions, conditions change seldom, damage/theft is possible	☐ **mild, never changing** mild conditions, conditions change never, damage/theft is not likely
Mobility Choose whether the sensor/device/machine is physically moving during operation.	☐ **always** moving most of the time	☐ **sometimes** moving sometimes	☐ **never** never moving
Energy			
Power source Choose the type of power supply of the sensor/device/machine.	☐ **battery** battery-powered		☐ **hardwired** wired power supply
Usage Choose the power consumption behavior of the sensor/device/machine.	☐ **low** rarely on, passive mode/sleep mechanisms used	☐ **medium** power saving, passive mode/sleep mechanisms used	☐ **high** always on, no passive mode/sleep mechanisms used
Connectivity			
Mechanism Choose the physical mechanism used to convey any data from the sensor/device/machine.	☐ **wireless** spectrum, light, sound; e.g. radio frequency (e.g. RFID, Bluetooth Low Energy, Wi-Fi, ZigBee)	☐ **physical** pressure, mechanical, thermal; e. g. ulltrasonic communication (e.g. SoniTalk)	☐ **wired** electrical, fibre optic; e.g. Ethernet
Hardware			
Quality Choose the hardware-related quality level of the sensor/device/machine.	☐ **low** low cost sensors, commodity nature, not suitable for use in the application used	☐ **medium** state-of-the art quality, suitable for use in the application used	☐ **high** highly precise and robust hardware
Maintenance/calibration Choose the frequency of maintenance activities that are carried out on the sensor/device/machine.	☐ **never** never serviced	☐ **sporadic** sporadic serviced (e.g. in case of failures)	☐ **regulary** regulary serviced
Ease of replacement/repair Choose the level of difficulty to repair or replace the sensor/device/machine in case of faulty or malfunctioning behavior.	☐ **hard** unsupported, no experience, no access	☐ **medium** supported, no experience, accessable	☐ **easy** supported, knowledge available, accessable
Constraints Choose the extent of resource constraints the sensor/device/machine suffers from.	☐ **severe limitations** low battery capacity, low-performance hardware, no backup power source	☐ **moderate limitations** state-of-the art hardware	☐ **no limitations** high battery capacity, high-performance hardware, backup power available
Data preprocessing			
Choose the level of data preprocessing that is carried out on the sensor/device/machine.	☐ **high** a lot of data preprocessing carried out	☐ **medium** a state-of-the art level of preprocessing carried out	☐ **low/no** no essential data preprocessing carried out

Figure C.18: Data provider property checklist.

Appendix D. Empirical Validation

Appendix D.1. Filled Data Provider Property Checklist

Properties		Choice		
General				
	Age Choose the age of the sensor/device/machine.	☑ **old** legacy technology, operating hours exceed average lifespan	☐ **middle** state-of-the art technology, operating hours lower average lifespan	☐ **new** new technology, few operating hours
	Provider type Choose whether the provided data are partially based on human input or completely created by the sensor/device/machine itself.	☑ **persons involved** (partial) manual data entry/dependence		☐ **only sensor/device/machine** no manual data entry/dependence
Location				
	Environment Choose the physical environment in which the sensor/device/machine operates.	☐ **harsh, dynamically changing** extreme conditions, conditions change often, likely to damage/theft	☑ **moderate, seldom changing** moderate conditions, conditions change seldom, damage/theft is possible	☐ **mild, never changing** mild conditions, conditions change never, damage/theft is not likely
	Mobility Choose whether the sensor/device/machine is physically moving during operation.	☑ **always** moving most of the time	☐ **sometimes** moving sometimes	☐ **never** never moving
Energy				
	Power source Choose the type of power supply of the sensor/device/machine.	☑ **battery** battery-powered		☐ **hardwired** wired power supply
	Usage Choose the power consumption behavior of the sensor/device/machine.	☐ **low** rarely on, passive mode/sleep mechanisms used	☑ **medium** power saving, passive mode/sleep mechanisms used	☐ **high** always on, no passive mode/sleep mechanisms used
Connectivity				
	Mechanism Choose the physical mechanism used to convey any data from the sensor/device/machine.	☑ **wireless** spectrum, light, sound; e.g. radio frequency (e.g. RFID, Bluetooth Low Energy, Wi-Fi, ZigBee)	☐ **physical** pressure, mechanical, thermal; e. g. ulltrasonic communication (e.g. SoniTalk)	☐ **wired** electrical, fibre optic; e.g. Ethernet
Hardware				
	Quality Choose the hardware-related quality level of the sensor/device/machine.	☑ **low** low cost sensors, commodity nature, not suitable for use in the application used	☐ **medium** state-of-the art quality, suitable for use in the application used	☐ **high** highly precise and robust hardware
	Maintenance/calibration Choose the frequency of maintenance activities that are carried out on the sensor/device/machine.	☑ **never** never serviced	☐ **sporadic** sporadic serviced (e.g. in case of failures)	☐ **regulary** regulary serviced
	Ease of replacement/repair Choose the level of difficulty to repair or replace the sensor/device/machine in case of faulty or malfunctioning behavior.	☐ **hard** unsupported, no experience, no access	☑ **medium** supported, no experience, accessable	☐ **easy** supported, knowledge available, accessable
	Constraints Choose the extent of resource constraints the sensor/device/machine suffers from.	☑ **severe limitations** low battery capacity, low-performance hardware, no backup power source	☐ **moderate limitations** state-of-the art hardware	☐ **no limitations** high battery capacity, high-performance hardware, backup power available
Data preprocessing				
	Choose the level of data preprocessing that is carried out on the sensor/device/machine.	☐ **high** a lot of data preprocessing carried out	☑ **medium** a state-of-the art level of preprocessing carried out	☐ **low/no** no essential data preprocessing carried out

Figure D.19: Filled data provider checklist for the wearable barcode scanner combined with a smartphone used in the empirical validation.

Appendix D.2. Data Quality Assessments per Expert

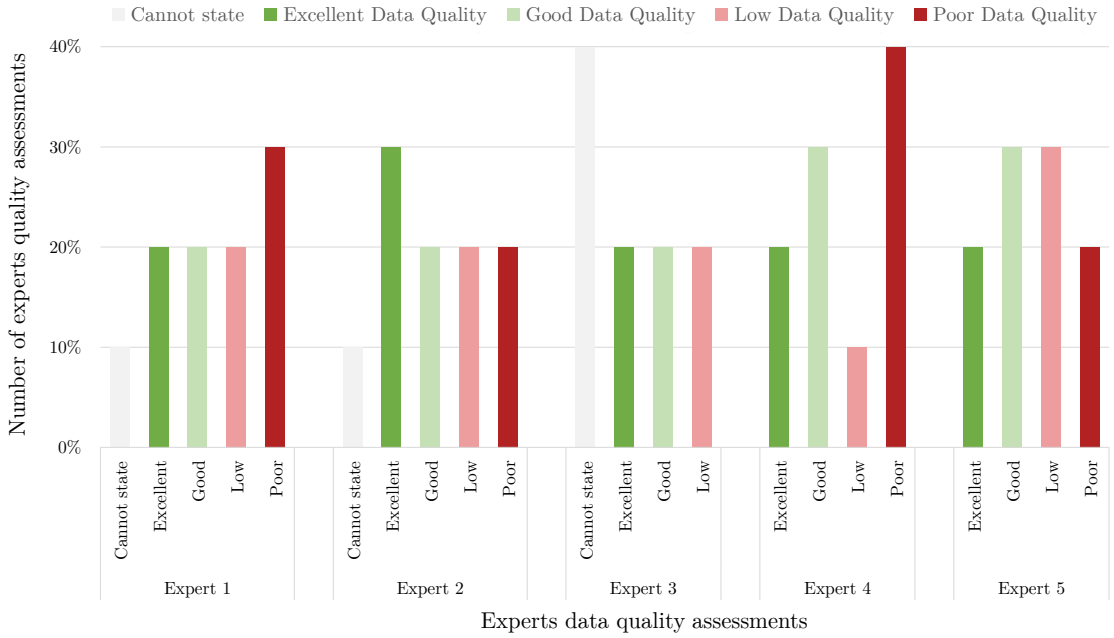

Figure D.20: Overview of the distribution of the data quality assessments for each expert.

Appendix D.3. Data Quality Assessments of the Experts per Data Source

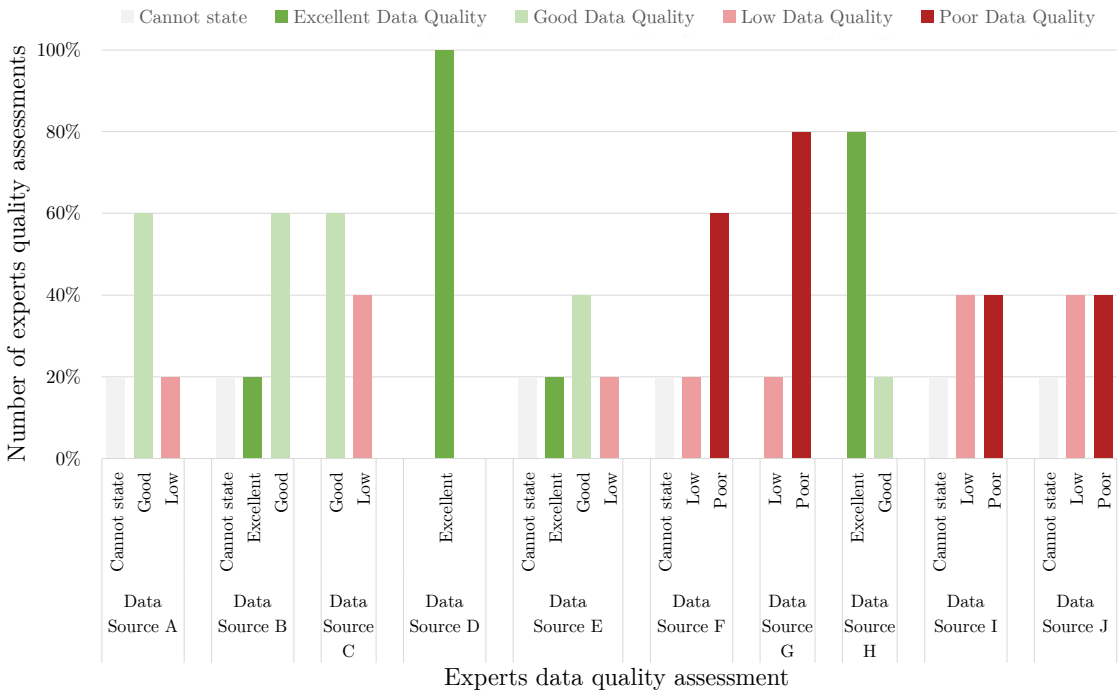

Figure D.21: Overview of the distribution of the data quality assessments of the experts for each data source.

Appendix D.4. Trust Scores (Data Source Assessment Approach)

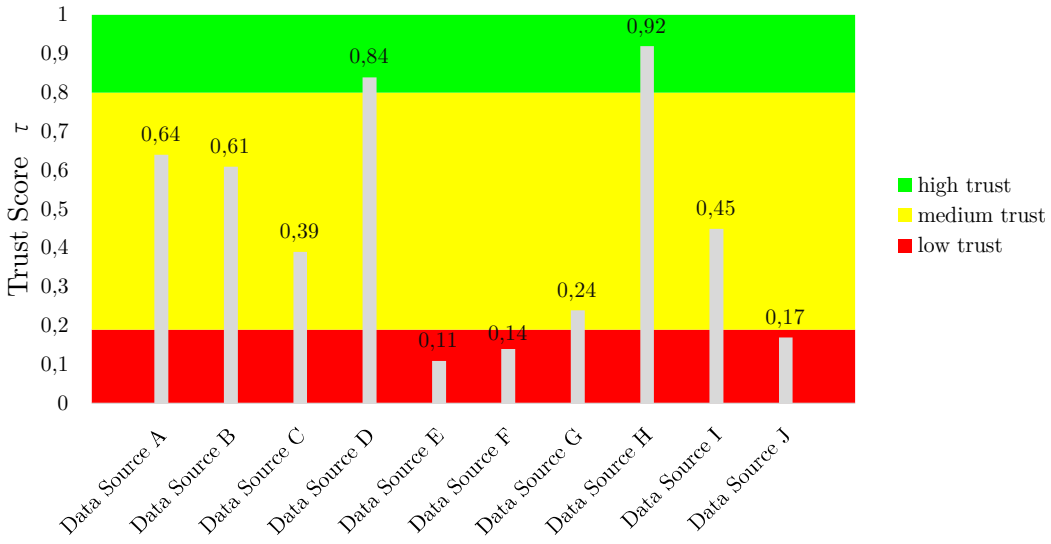

Figure D.22: Trust scores and trust categories for each data source based on the data source assessment approach.

References

[1] P. C. Evans, M. Annunziata, Industrial internet: Pushing the boundaries of minds and machines (2012).
URL https://www.ge.com/docs/chapters/Industrial_Internet.pdf

[2] H. Kagermann, Recommendations for implementing the strategic initiative industrie 4.0 (2013).

[3] X. E. Lee, Made in china 2025: A new era for chinese manufacturing, CKGSB Knowledge (2015).

[4] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, M. Gidlund, Industrial internet of things: Challenges, opportunities, and directions, IEEE Transactions on Industrial Informatics 14 (11) (2018) 4724–4734.

[5] H. D. Nguyen, K. P. Tran, X. Zeng, L. Koehl, P. Castagliola, P. BRUNIAUX, Industrial internet of things, big data, and artificial intelligence in the smart factory: a survey and perspective, in: ISSAT International Conference on Data Science in Business, Finance and Industry, Da Nang, Vietnam, 2019, pp. 72–76.
URL https://hal.archives-ouvertes.fr/hal-02268119

[6] H. Boyes, B. Hallaq, J. Cunningham, T. Watson, The industrial internet of things (iiot): An analysis framework, Computers in Industry 101 (2018) 1–12. doi:10.1016/j.compind.2018.04.015.

[7] Y. Yu, R. Chen, H. Li, Y. Li, A. Tian, Toward data security in edge intelligent iiot, IEEE Network 33 (5) (2019) 20–26. doi:10.1109/MNET.001.1800507.

[8] Grand View Research, Industrial internet of things market size, share & trends analysis report by component, by end use (manufacturing, energy & power, oil & gas, healthcare, logistics & transport, agriculture), and segment forecasts, 2019 - 2025 (01.06.2019).
URL https://www.grandviewresearch.com/industry-analysis/industrial-internet-of-things-iiot-market

[9] Industrial Internet Consortium, The industrial internet of things volume t3: Analytics framework: Iic:pub:t3:v1.00:pb:20171023 (2017).

[10] P. Lade, R. Ghosh, S. Srinivasan, Manufacturing analytics and industrial internet of things, IEEE Intelligent Systems 32 (3) (2017) 74–79. doi:10.1109/MIS.2017.49.

[11] C. Keßler, R. T. A. de Groot, Trust as a proxy measure for the quality of volunteered geographic information in the case of openstreetmap, in: D. Vandenbroucke, B. Bucher, J. Crompvoets (Eds.), Geographic Information Science at the Heart of Europe, Springer International Publishing, Cham, 2013, pp. 21–37.

[12] J. Byabazaire, G. O'Hare, D. Delaney, Using trust as a measure to derive data quality in data shared iot deployments, in: 2020 29th International Conference on Computer Communications and Networks (ICCCN), 2020, pp. 1–9.

[13] H. u. Rahman, G. Wang, M. Z. Alam Bhuiyan, J. Chen, Trustworthy data collection for cyber systems: A taxonomy and future directions, in: G. Wang, A. El Saddik, X. Lai, G. Martinez Perez, K.-K. R. Choo (Eds.), Smart City and Informatization, Vol. 1122 of Communications in Computer and Information Science, Springer Singapore, Singapore, 2019, pp. 152–164.

[14] N. Haron, J. Jaafar, I. A. Aziz, M. H. Hassan, 2017 IEEE Conference on Big Data and Analytics: 16th-17th November 2017, Riverside Majestic Hotel, Kuching, Malaysia, IEEE, Piscataway, NJ, 2017.

[15] L.-A. Tang, X. Yu, S. Kim, Q. Gu, J. Han, A. Leung, T. La Porta, Trustworthiness analysis of sensor data in cyber-physical systems, Journal of Computer and System Sciences 79 (3) (2013) 383–401.

[16] S. Sicari, C. Cappiello, F. de Pellegrini, D. Miorandi, A. Coen-Porisini, A security-and quality-aware system architecture for internet of things, Information Systems Frontiers 18 (4) (2016) 665–677. doi:10.1007/s10796-014-9538-x.
[17] H.-S. Lim, Y.-S. Moon, E. Bertino, Provenance-based trustworthiness assessment in sensor networks, in: D. Zeinalipour, W.-C. Lee (Eds.), Proceedings of the Seventh International Workshop on Data Management for Sensor Networks - DMSN '10, ACM Press, New York, New York, USA, 2010, p. 2. doi:10.1145/1858158.1858162.
[18] A. Satter, N. Ibtehaz, A regression based sensor data prediction technique to analyze data trustworthiness in cyber-physical system, International Journal of Information Engineering and Electronic Business 10 (3) (2018) 15–22. doi:10.5815/ijieeb.2018.03.03.
[19] S. Zhao, J. Wen, S. Mumtaz, S. Garg, B. J. Choi, Spatially coupled codes via partial and recursive superposition for industrial iot with high trustworthiness, IEEE Transactions on Industrial Informatics 16 (9) (2020) 6143–6153. doi:10.1109/TII.2020.2965952.
[20] C. Dai, D. Lin, E. Bertino, M. Kantarcioglu, An approach to evaluate data trustworthiness based on data provenance, in: W. Jonker, M. Petković (Eds.), Secure Data Management, Vol. 5159 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 82–98.
[21] E. Bertino, Data trustworthiness—approaches and research challenges, in: J. Garcia-Alfaro, J. Herrera-Joancomartí, E. Lupu, J. Posegga, A. Aldini, F. Martinelli, N. Suri (Eds.), Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance, Vol. 8872 of Lecture Notes in Computer Science, Springer International Publishing, Cham, 2015, pp. 17–25.
[22] T. Sharma, M. Fragkoulis, S. Rizou, M. Bruntink, D. Spinellis, Smelly relations: Measuring and understanding database schema quality, in: 2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP), 2018, pp. 55–64.
[23] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, Data quality in internet of things: A state-of-the-art survey, Journal of Network and Computer Applications 73 (2016) 57–81.
[24] C. C. Aggarwal, N. Ashish, A. Sheth, The internet of things: A survey from the data-centric perspective, in: C. C. Aggarwal (Ed.), Managing and Mining Sensor Data, Springer US, Boston, MA, 2013, pp. 383–428.
[25] H. Y. Teh, A. W. Kempa-Liehr, K. I.-K. Wang, Sensor data quality: a systematic review, Journal of Big Data 7 (2020). doi:10.1186/s40537-020-0285-1.
[26] D.-Y. Kim, Y.-S. Jeong, S. Kim, Data-filtering system to avoid total data distortion in iot networking, Symmetry 9 (16) (2017). doi:10.3390/sym9010016.
[27] U. Wetzker, I. Splitt, M. Zimmerling, C. A. Boano, K. Romer, Troubleshooting wireless coexistence problems in the industrial internet of things, in: 2016 IEEE Intl Conference, 2016, p. 98.
[28] C. Alexakos, A. Komninos, C. Anagnostopoulos, G. Kalogeras, A. Savvopoulos, A. Kalogeras, Building an industrial iot infrastructure with open source software for smart energy, in: 2019 First International Conference on Societal Automation (SA), IEEE, 04.09.2019 - 06.09.2019, pp. 1–8.
[29] M. H. u. Rehman, E. Ahmed, I. Yaqoob, I. A. T. Hashem, M. Imran, S. Ahmad, Big data analytics in industrial iot using a concentric computing model, IEEE Communications Magazine 56 (2) (2018) 37–43. doi:10.1109/MCOM.2018.1700632.
[30] W. Z. Khan, M. H. Rehman, H. M. Zangoti, M. K. Afzal, N. Armi, K. Salah, Industrial internet of things: Recent advances, enabling technologies and open challenges, Computers & Electrical Engineering 81 (2020) 106522.
[31] M. H. u. Rehman, I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, C. Perera, The role of big data analytics in industrial internet of things, Future Generation Computer Systems 99 (2019) 247–259.
[32] IoT Analytics GmbH, Industrial ai market report 2020-2025 (2019).
URL https://iot-analytics.com/product/industrial-ai-market-report-2020-2025/
[33] P. Kamat, R. Sugandhi, Anomaly detection for predictive maintenance in industry 4.0- a survey, E3S Web of Conferences 170 (2020) 02007. doi:10.1051/e3sconf/202017002007.
[34] H. M. Hashemian, W. C. Bean, State-of-the-art predictive maintenance techniques*, IEEE Transactions on Instrumentation and Measurement 60 (10) (2011) 3480–3492. doi:10.1109/TIM.2009.2036347.
[35] F. Tao, Q. Qi, A. Liu, A. Kusiak, Data-driven smart manufacturing, Journal of Manufacturing Systems 48 (2018) 157–169. doi:10.1016/j.jmsy.2018.01.006.
[36] Industrial Internet Consortium, The industrial internet of things volume g1: Reference architecture: Version 1.9 (19.06.2019).
[37] G. Veneri, A. Capasso, Hands-on industrial internet of things: Create a powerful industrial IoT infrastructure using Industry 4.0, Packt Publishing, 2018.
[38] R. Y. Wang, V. C. Storey, C. P. Firth, A framework for analysis of data quality research, IEEE Transactions on Knowledge and Data Engineering 7 (4) (1995) 623–640. doi:10.1109/69.404034.
[39] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems 12 (4) (1996) 5–33.
[40] C. Cichy, S. Rass, An overview of data quality frameworks, IEEE Access 7 (2019) 24634–24648.
[41] A. Caro, C. Calero, I. Caballero, M. Piattini, A proposal for a set of attributes relevant for web portal data quality, Software Quality Journal 16 (4) (2008) 513–542. doi:10.1007/s11219-008-9046-7.
[42] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, A. Gómez-Pérez, A comprehensive quality model for linked data, Semantic Web 9 (1) (2018) 3–24.
[43] J. Merino, I. Caballero, B. Rivas, M. Serrano, M. Piattini, A data quality in use model for big data, Future Generation Computer Systems 63 (2016) 123–130.
[44] L. L. Pipino, Y. W. Lee, R. Y. Wang, Data quality assessment, Communications of the ACM 45 (4) (2002) 211–218.
[45] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM

computing surveys (CSUR) 41 (3) (2009) 1–52.
[46] ISO/IEC, Iso/iec 25012:2008 software engineering - software product quality requirements and evaluation (square) - data quality model: Iso/iec (2008).
[47] H. Foidl, M. Felderer, S. Biffl, Technical debt in data-intensive software systems, in: 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2019, pp. 338–341.
[48] J. Gao, C. Xie, C. Tao, Big data validation and quality assurance – issuses, challenges, and needs, in: 2016 IEEE Symposium on Service-Oriented, 2016, pp. 433–441. doi:10.1109/SOSE.2016.63.
[49] V. Gudivada, A. Apon, J. Ding, Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations, International Journal on Advances in Software 10 (1) (2017) 1–20.
[50] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data science journal 14 (2015).
[51] H. Tao, M. Z. A. Bhuiyan, M. A. Rahman, T. Wang, J. Wu, S. Q. Salih, Y. Li, T. Hayajneh, Trustdata: Trustworthy and secured data collection for event detection in industrial cyber-physical system, IEEE Transactions on Industrial Informatics 16 (5) (2020) 3311–3321. doi:10.1109/TII.2019.2950192.
[52] P. J. H. Daas, S. J. L. Ossen, J. Arends-Tóth, Framework of quality assurance for administrative data sources, in: 57th World Statistics Congress ISI, Vol. 1622, 2009.
[53] D. Dufty, H. Bérard, S. Lefranc, M. Signore, A suggested framework for the quality of big data, UNECE big data quality task team (2014).
[54] U. Milošević, B. van Nuffelen, P. Massey, D3.2 data source assessment methodology v2.0: 645886 — yds — h2020-inso-2014-2015/h2020-inso-2014 (2016).
[55] M. Solar, G. Concha, L. Meijueiro, A model to assess open government data in public agencies, in: International Conference on Electronic Government, 2012, pp. 210–221.
[56] H. Foidl, M. Felderer, Risk-based data validation in machine learning-based software systems, in: proceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation, 2019, pp. 13–18.
[57] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering (2007).
[58] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: M. Shepperd, T. Hall, I. Myrtveit (Eds.), Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14, ACM Press, New York, New York, USA, 2014, pp. 1–10.
[59] M. Jarke, M. A. Jeusfeld, C. Quix, P. Vassiliadis, Architecture and quality in data warehouses: An extended repository approach, Information Systems 24 (3) (1999) 229–253.
[60] A. Nazabal, C. K. I. Williams, G. Colavizza, C. R. Smith, A. Williams, Data engineering for data analytics: A classification of the issues, and case studies (2020).
[61] O. Romero, R. Wrembel, Data engineering for data science: Two sides of the same coin, in: International Conference on Big Data Analytics and Knowledge Discovery, 2020, pp. 157–166.
[62] L. Cao, Data science: Profession and education, IEEE Intelligent Systems 34 (5) (2019) 35–44.
[63] A. Alvaro, E. S. Almeida, S. L. Meira, Quality attributes for a component quality model, 10th WCOP/19th ECCOP, Glasgow, Scotland (2005) 31–37.
[64] L. Ehrlinger, W. Wöß, Automated schema quality measurement in large-scale information systems, in: International Workshop on Data Quality and Trust in Big Data, 2018, pp. 16–31.
[65] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, P. Hitzler, Quality assessment methodologies for linked open data, Submitted to Semantic Web Journal 1 (2013) 1–5.
[66] T. Punter, Using checklists to evaluate software product quality, in: Proceedings of the 8th European Conference on Software Cost Estimation (ESCOM), 1997, pp. 143–150.
[67] J. R. Lourenço, B. Cabral, P. Carreiro, M. Vieira, J. Bernardino, Choosing the right nosql database for the job: a quality attribute evaluation, Journal of Big Data 2 (1) (2015) 18.
[68] N. Zubair, N. A, K. Hebbar, Y. Simmhan, Characterizing iot data and its quality for use (2019).
[69] ISO, Iso 8000-61:2016: Data quality - part 61: Data quality management: Process reference model (2016).
URL https://www.iso.org/standard/63086.html
[70] S. Wagner, A. Goeb, L. Heinemann, M. Kläs, C. Lampasona, K. Lochmann, A. Mayr, R. Plösch, A. Seidl, J. Streit, A. Trendowicz, Operationalised product quality models and assessment: The quamoco approach, Information and Software Technology 62 (2015) 101–123.
[71] A. F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data, Communication methods and measures 1 (1) (2007) 77–89.
[72] K. Krippendorff, Computing krippendorff's alpha-reliability (2011).
[73] K. Krippendorff, Content analysis: an introduction to its methodology sage, Thousand Oaks, CA (2004).
[74] R. Taylor, Interpretation of the correlation coefficient: a basic review, Journal of diagnostic medical sonography 6 (1) (1990) 35–39.
[75] M. Al-Barak, R. Bahsoon, Database design debts through examining schema evolution, in: 2016 IEEE 8th International Workshop on Managing Technical Debt (MTD), 2016, pp. 17–23.
[76] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empirical Software Engineering 14 (2) (2009) 131–164. doi:10.1007/s10664-008-9102-8.
[77] K. Petersen, C. Gencel, Worldviews, research methods, and their relationship to validity in empirical software engineering research, in: 2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement, IEEE, 23.10.2013 - 26.10.2013, pp. 81–89.

An Approach for Assessing Industrial IoT Data Sources to Determine their Data Trustworthiness

Harald Foidl[a], Michael Felderer[a,b]

[a] *University of Innsbruck, Austria*
[b] *Blekinge Institute of Technology, Sweden*

Preprint submitted to Internet of Things *March 9, 2022*