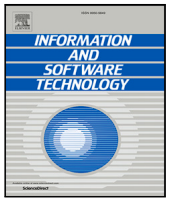
Information and Software Technology 167 (2024) 107378

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



BIGOWL4DQ: Ontology-driven approach for Big Data quality meta-modelling, selection and reasoning

Cristóbal Barba-González [a,*], Ismael Caballero [b], Ángel Jesús Varela-Vaca [c], José A. Cruz-Lemus [b], María Teresa Gómez-López [c], Ismael Navas-Delgado [a]

[a] KHAOS Research group, ITIS Software, Universidad de Málaga, Málaga 29071, Spain
[b] Institute of Technologies and Information Systems, University of Castilla-La Mancha, Ciudad Real, Spain
[c] IDEA Research group, Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Sevilla, Spain

ARTICLE INFO

Keywords:
Data quality evaluation and measurement
Data quality information model
Big Data
Ontology
Decision model and notation

ABSTRACT

Context: Data quality should be at the core of many Artificial Intelligence initiatives from the very first moment in which data is required for a successful analysis. Measurement and evaluation of the level of quality are crucial to determining whether data can be used for the tasks at hand. Conscientious of this importance, industry and academia have proposed several data quality measurements and assessment frameworks over the last two decades. Unfortunately, there is no common and shared vocabulary for data quality terms. Thus, it is difficult and time-consuming to integrate data quality analysis within a (Big) Data workflow for performing Artificial Intelligence tasks. One of the main reasons is that, except for a reduced number of proposals, the presented vocabularies are neither machine-readable nor processable, needing human processing to be incorporated.

Objective: This paper proposes a unified data quality measurement and assessment information model. This model can be used in different environments and contexts to describe data quality measurement and evaluation concerns.

Method: The model has been developed as an ontology to make it interoperable and machine-readable. For better interoperability and applicability, this ontology, BIGOWL4DQ, has been developed as an extension of a previously developed ontology for describing knowledge management in Big Data analytics.

Conclusions: This extended ontology provides a data quality measurement and assessment framework required when designing Artificial Intelligence workflows and integrated reasoning capacities. Thus, BIGOWL4DQ can be used to describe Big Data analysis and assess the data quality before the analysis.

Result: Our proposal has been validated with two use cases. First, the semantic proposal has been assessed using an academic use case. And second, a real-world case study within an Artificial Intelligence workflow has been conducted to endorse our work.

1. Introduction

Nowadays, the rising volume and the heterogeneity of the types of data made necessary the development of Big Data technologies that facilitate data preparation and Artificial Intelligence (AI) analysis [1]. It is also crucial to define a context where the data quality terms must be specified [2]. To achieve the data quality requirements, data preparation must be performed, but it is not a recipe applied to a dataset; it must be closely related to the specific levels of data quality required for the intended use, depending on the organisation and the moment. Frequently, data scientists consider the actions needed for data quality management only locally for their problem [3]. With broader visions of data quality across data for several issues, the efforts achieved to increase data quality can be reduced and shared [4]. For this reason, it is necessary to use a Data Quality Assessment Framework, such as SparkDQ [5], Drunken Data Quality [6], Deequ [7], or Apache Griffin [8], that enables sharing of data quality assessment and improvement results as part of a context-aware Data Preparation Process. However, as we propose in this work, the Data Quality Assessment frameworks do not assess their data quality rules during the design phase to find flaws or inconsistencies before applying them.

The wide range of contexts in which the same data can potentially be used makes it necessary to define a standard view of the quality

* Corresponding author.
E-mail addresses: cbarba@uma.es (C. Barba-González), ismael.caballero@uclm.es (I. Caballero), ajvarela@us.es (Á.J. Varela-Vaca), joseantonio.cruz@uclm.es (J.A. Cruz-Lemus), maytegomez@us.es (M.T. Gómez-López), ismael@uma.es (I. Navas-Delgado).

https://doi.org/10.1016/j.infsof.2023.107378
Received 31 March 2023; Received in revised form 17 November 2023; Accepted 26 November 2023
Available online 27 November 2023

of data repositories. This is even more necessary due to the wide variety of sources required to feed data repositories, derived from the necessity to integrate various data requirements and formats. Typically, the main activities of the data preparation process are exploration, structure, cleaning, and shaping [9]. One reason why this process is widely recognised as the most time-consuming task for data analysis is because of the lack of shared information throughout the process about the dataset, including the data's structure and semantics and the data quality levels [10]. In this sense, it is necessary to recall that, without adequate mechanisms, only humans can process raw data, typically coming from various sources with different structures and representation/implementation but with the same meaning.

Consequently, actions related to data preparation largely depend on the understanding that humans can extract from the data, including levels of data quality. Thus, it can be said that the mechanism of data preparation depends on the knowledge humans have acquired from observing and using the data. The data preparation process will be benefited from optimising by automating some of these efforts [11].

Business rules are related to data quality in two leastwise ways. First, they let to guide the company's decisions in its daily operations. Second, they permit the auditing of data produced by existing processes for compliance with external regulations and internal business policies and goals. Thus, business rules offer new perspectives to improve data quality which is essential to have high-quality data to ensure that the results of data analysis are reliable and valuable [12,13].

In this paper, we focus on automating the production and sharing of the data quality assessment results, integrating these results with the metadata of the Big Data workflows, to make the process of applying Artificial Intelligence techniques more efficient. To achieve this goal, we start with *BIG data analytics OWL ontology* (*BIGOWL*) [2] and the ontology approach to support knowledge management in Big Data analytics. BIGOWL includes concepts for representing Big Data analytics workflows, including different dimensions: components, workflow definition, and domain knowledge. Details are described in a functional classification that helps them be found when building workflows. However, BIGOWL does not include relevant data management dimensions, such as cybersecurity or data quality. To fill this void, we propose to add the data quality dimension since, as previously said, data quality management must be closed before applying AI solutions (i.e., Machine Learning, Deep Learning, Optimisation).

The World Wide Web Consortium (W3C)[1] proposed the Data Quality Vocabulary (DQ-Vocab),[2] which provides a metadata model for data quality. However, DQ-Vocab or other existing ontologies [14–18] do not cover the relationship between data quality and data use. Most data quality assessment methodologies [19] directly point to the need to state and validate the corresponding business rules to support the measurement and assessment of data quality. In that regard, it is interesting to note that the Decision Model and Notation (DMN) [20] supports the modelling and evaluation of business rules, described using an expressive language called S-FEEL (Simplified Friendly Enough Expression Language) that combines fundamental expressions [21,22]. Using DMN tables, there are frameworks [23–25] for modelling and evaluation of data quality rules that follow the DMN standard. *DMN4DQ* a framework that uses the rules to represent the user requirements for the quality of the data in the context of use to generate a recommendation on the usability of the data is presented in [25]. This recommendation is based on evaluating the quality of the data for this context of use. However, in this framework, the rules are not evaluated to identify flaws or inconsistencies before they have been executed. For this reason, we propose the creation of the *BIGOWL4DQ* ontology, whose main aim is representing and consolidating data quality knowledge for Big Data analytics.

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

[2] https://www.w3.org/TR/vocab-dqv/.

BIGOWL4DQ extends BIGOWL with data quality dimensions, which provide mechanisms for improving the reasoning capacity and facilitate the incorporation of data quality tasks in Big Data workflows. Thus, our threefold objectives are (1) to represent actionable knowledge to automate the process of better-supporting data preparation; (2) to reduce the complexity of defining the main concepts of Data Quality management within a context-aware Data Preparation Process; and (3) to assess the correctness and completeness of the data quality rules. The challenge of guaranteeing the accuracy of data quality rules, specifically identifying inconsistent or incomplete rules. For example, DMN decision tables are raised by using DMN decision tables as a specification vehicle for crucial business decisions [26]. Reasoning employing BIGOWL4DQ annotations helps identify errors in the rules at the specification stage, which may help avoid costly flaws later on during the design and execution of business processes. We complement the OWL 2 axioms with SWRL (Semantic Web Rule Language) [27] rules for the reasoning process.

1.1. Contributions

The main contributions of this study are:

- The proposed ontology, BIGOWL4DQ, which has been designed and implemented for a comprehensive solution that integrates and appropriately relates the measurement, assessment, and creation of a suggestion for using data through business rules.
- A semantic approach has been modelled and implemented to annotate all the meta-data involved from business rules for defining the main concepts of Data Quality measurement, management, and Data Quality assessment.
- BIGOWL4DQ has been designed to include SWRL rules for evaluating the correctness and completeness of business rules for data quality in measurement, assessment, and usability decisions.
- The semantic model is evaluated in two contexts, one from an academic use case and the other from a practical use case of smart farms related to soil sensor networks.

1.2. Structure of this work

The rest of the paper is structured as follows: Section 2 introduces the background needed to understand the proposal and the related work; Section 3 details the semantic method required to assess the data quality using the case study from [25] to illustrate the proposal; Section 4 presents the application of our proposal to the case study; in Section 5 we point out the main threads to validity of this work; Discussion and conclusions are presented in Section 6.

2. Foundations

This section introduces the theoretical foundations required to develop the investigation introduced in this manuscript. There are two important groups: one aimed at introducing concepts related to the representation of ontologies and the other dedicated to the field of data quality measurement. A review of the state-of-the-art is also given to highlight the main differences between the related works and the proposed approach.

2.1. Background concepts on ontology and representation of knowledge

An *ontology* offers a formal representation of the real world [28]. It provides properties of each idea, which describe numerous aspects and qualities of concepts (classes of concepts), limitations on properties, and an explicit explanation of concepts in a domain of discourse. A knowledge base comprises an ontology and a collection of unique instances of classes, and it provides services to encourage interoperability across various heterogeneous systems and databases [29].

OWL is an ontology language based on decidable subsets of first-order logic: Description Logics (DL). As for DL in OWL, we can differentiate between the T-Box (concepts, relationships and constraints) [30] and the A-Box (relations between individuals and concepts). OWL as a Knowledge Representation Model enables automated inference processes to extract implicit knowledge from a knowledge base. The use of OWL also enables homogenisation in the description process using the explicit knowledge represented, easing the integration of different datasets. The OWL 2 Web Ontology Language, informally OWL 2, is the latest W3C Recommendation for ontologies. There are some OWL 2 editors, but the most well-known is Protégé.[3] It can be used as a standalone application or a Web-based tool, enabling the team development of OWL ontologies.

Semantic Web Rule Language(SWRL)[4] is a W3C proposal to combine OWL 2 with the Rule Markup Language (RuleML). However, it only applies to the OWL DL subset and the Unary/Binary Datalog RuleML sublanguages. SWRL can be evaluated using existing OWL 2 Reasoners. Thus, this language increases the reasoning capabilities of OWL 2 with production rules.

OWL-based ontologies are given procedural knowledge by the Semantic Web Rule Language (SWLR), which makes up for some of the shortcomings of ontology inference, notably in recognising semantic links between instances [31]. SWRL includes a high-level abstract syntax for Horn-like rules.[5] Model-theoretic semantics provides the formal meaning of OWL 2 ontologies, including rules written in this abstract syntax. An OWL 2 ontology is made up of a list of rules and facts. Each rule has an antecedent and a consequent, where if the antecedent is true, then the consequent must be satisfied. Facts are rules without an antecedent.

SWRL provides the typical logic expression "$Antecedent \Rightarrow Consequent$" to represent semantic rules. The antecedent of the pair (rule body) and the consequent (rule head) can be conjunctions of one or more atoms written as "$atom_1 \wedge atom_2 \wedge \cdots \wedge atom_n$". Each atom has one or more associated parameters, which are denoted by a question mark and a variable (e.g., $?x$). The typical uses of SWRL include the transfer of characteristics and inferring the existence of new individuals [32].

Both *TBox* and *ABox* are used in the reasoning processes of OWL 2. As in the case of ontology editors, many tools are available to perform querying, inference, and reasoning tasks on ontologies. *Reasoners* aim at finding implicit knowledge in the ontologies by applying a set of reasoning mechanisms, such as:

- *Satisfiability*: Check whether an ontology class can be instantiated.
- Subsumption: Check whether any ontology constraint is implied by the rest of the constraints.
- *Classification*:
  - Calculate the set of subclasses of each class explicitly, i.e. determine whether there are implicit subclasses.
  - Calculate whether two concepts are synonymous.
  - Calculate the most specific class to which an instance belongs.
- *Consistency*: Check whether the instances defined within an ontology satisfy all the restrictions.

As stated before, in this work, we extend with data quality concepts the ontology *BIGOWL*, whose main aim is to provide a broad vocabulary of terms related to Big Data analytics processes, including their components and how they are integrated, from data sources to analytics visualisation. This ontology was created using the OWL 2 ontology language, which uses classes to represent concepts and data attributes or object properties to describe relations. BIGOWL includes 184 classes, 66 individuals, 488 axioms, 16 object properties, and 20 data properties (individual attributes). In addition, BIGOWL also supports reasoning thanks to the formulation of semantic rules to deduce new information from existing knowledge. These rules are formulated in SWRL and are used to perform semantic reasoning jobs mainly devoted to checking the consistency of workflows. Reasoning tasks are evaluated in this work using Pellet [33], an open-source Java based OWL 2 reasoner.

[3] https://protege.stanford.edu/.
[4] https://www.w3.org/Submission/SWRL/.
[5] https://www.w3.org/2005/rules/wg/wiki/Horn_Rules_Semantics.html.

Table 1
Data quality dimensions by Wang et al. [39]. Data quality dimensions are the criteria for defining data quality category. Data quality dimensions indicate concepts, principles, and procedures for describing, measuring, analysing, and improving the critical aspects of data. Each row of the table represents the category and its dimensions.

Data quality category	Data quality dimension
Intrinsic	Accuracy, objectivity, believability, reputation
Accessibility	Access, security
Contextual	Relevancy, value-added, timeliness, completeness, amount of data
Representational	Interpretability, ease of understanding, concise representation, and consistent representation

Table 2
Inherent data quality characteristics from ISO 25012 [40]. A generic data quality model that applies to structured data stored in an information system is defined by the ISO 25012 standard. It establishes five basic quality dimensions common to any standard: accuracy, completeness, consistency, credibility and currentness.

Inherent data quality characteristics	Definition
Accuracy	The degree to which data have attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.
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.
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.
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.
Currentness	The degree to which data has attributes that are of the right age in a specific context of use

2.2. Data quality background

Data quality is generally understood as *fitness for use* [34]. There are frameworks and proposals [5–8,23–25] for assessing data quality. They usually focus on satisfying data quality as the fundamental guarantee for data-based research, decision-making, and service [35–37]. However, to judge data quality in a given context [38], one or more criteria are necessary to evaluate or assess the quality. These criteria are commonly known as *data quality dimensions* in the scientific literature [39] or *data quality characteristics* in the context of ISO standards [40]. The set of various data quality characteristics that are useable and eligible (e.g., to describe the most representative user data quality requirements) is called *data quality model*. Over the years, several authors and practitioners have provided their own data quality model adapted to their specific context of use. However, two of the most widely used and referenced are the one proposed by [39] (see Table 1) or the one proposed by [40] (see Table 2).

It is necessary to deploy a specific *measurement method* to determine whether a dataset has an adequate level of quality for an exact data quality dimension or characteristic representing a particular data quality requirement. The measurement method typically depends on the technology in which the data set is implemented [25,41]. Furthermore, to calculate the measurement, a specific set of business rules must be

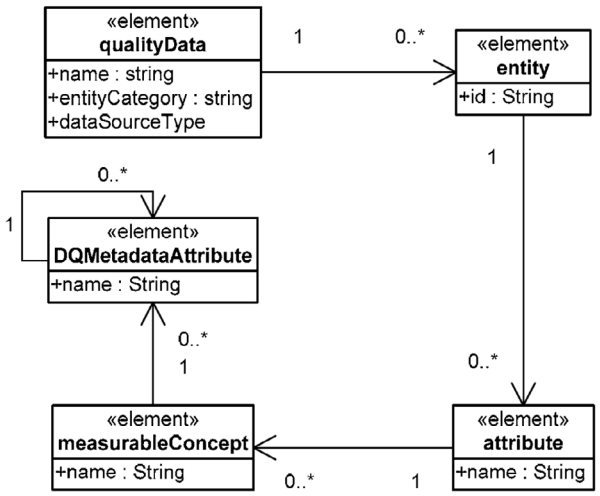


Fig. 1. Data Quality Measurement Information Model (DQMIM) [44] the referred terms by following ISO/IEC 15939 as a basis. These terms refer to the data quality employing concepts as an entity, attribute, and measurable concept.

conveniently selected and grouped for all data quality chosen characteristics [42]. The measurement result, typically calculated as the ratio of data records that do not violate the stated business rules and the total number of records, must then be compared with the specific *threshold values* representing the organisation's risk appetite for the task at hand. If the comparison is unfavourable, data quality analysts must decide on discarding or cleaning the data. Besides, depending on how bad the achieved measurement is obtained and the viability of the operation or the cost of cleaning the data is disproportionate, it may be easier to recapture or refuse to use the data. Thus, depending on the violations of business rules, specific actions must be designed and executed for data cleaning [43].

One of the most challenging works in the literature has been to represent the related knowledge that an organisation can manage about the levels of data quality of their data repositories. Apart from the specific sets of concepts provided in ISO/IEC 25012, ISO/IEC 25024, at this point, we would like to highlight two seminal works: DQMIM [44] and Vocab-DQ.[6] DQMIM, which stands for *Data Quality Measurement Information Model* (DQMIM), is based on the ISO/IEC 15939 standard [45] and provides a set of terms related to the measurement of software quality. These terms were adapted for the data quality, such as entity, attribute, and measurable concept (see Fig. 1). The main limitation of this model concerning the proposal of this work lies in the fact that, although it is sufficiently complete for a high level of abstraction, it does not allow for the incorporation of concrete details of the implementation of the measures.

In 2016, the W3C Working Group published *Vocab-DQ*, a Data Quality Vocabulary, to gather and align some previous work in the area with the ISO/IEC 25012 standard [40]. Their main aim was to provide a framework in which the quality of the datasets could be adequately described (see Fig. 2). Vocab-DQ is one of the most important contributions to representing data quality concepts using semantic web technologies. However, this vocabulary does not have the semantic capacity to express the relationship between the measurement and evaluation concepts described in DMN4DQ framework.

2.3. Rule modelling language for data quality description

The definition of data quality requirements through a set of rules facilitates modelling and a later evaluation with the tools that support the existing rule-modelling languages. Various standards support the modelling, description, and evaluation of business rules, such as the following:

- SBVR[7] defines the vocabularies and rules required to communicate organisations and software tools due to the definition of common elements. SBVR has a sound theoretical foundation of formal logic: it is based on first-order predicate logic with extensions into modal logic, i.e., some deontic forms for expressing obligation and prohibition and alethic forms for expressing necessities and possibilities [46].
- Semantic Web Rule Language (SWRL)[8] combines the OWL DL and OWL Lite sublanguages. SWRL includes a high-level abstract syntax for Horn-like rules. Additionally, model-theoretic semantics is given to provide the formal meaning for OWL ontologies, including rules in this abstract syntax. An OWL ontology is made up of a list of rules and facts. Each rule has the antecedent, and consequent, where if the antecedent is true, the consequent must be satisfied. Facts are rules without an antecedent.
- Production Rule Representation (PRR)[9] is an OMG standard to provide a vendor-neutral rule-model representation in UML for production rules. The standard provides a set of metamodels and OCL restrictions to define production rules enabling the afterwards transformation of those models to any rule engine.
- Decision Model and Notation (DMN)[10] is an OMG standard to obtain and represent decision models through a declarative description of the form 'if then' [47]. It is a standard notation for capturing decision logic that can be used in general business applications. DMN facilitates the modelling of repeatable decisions according to the necessities and is supported by a set of engines, such as Camunda - DMN Engine[11] or Drools - DMN.[12] DMN has also been used to represent data quality rules and validated in real datasets [20,25].
- RuleML (Rule Markup Language)[13] represents a family of languages for the specification of rules in the web context. RuleML can be used as a bridge between other OMG languages, such as SWRL, SBVR or PRR. There exist translators between RuleML and DMN languages, such as [48].

2.4. Related work

The use of semantics in the area of data quality management can be summarised in the following major uses [49]:

- **Collaborative representation and use of quality-relevant knowledge**. The semantics facilitates the definition of the data requirements in a structured and shareable way. Besides, semantics enables a machine-processable data format that is also readable by humans. This representation focuses on defining vocabularies [50], data dependencies [51], or rules [52].
- **Automated identification of conflicting data requirements into harmonised data**. For example, [53] focuses on using semantics to manage incomplete terminologies, irrelevant terms, outliers, missing values, data categorisation, and duplicated terms to make data standardisation in the medical data context.

[6] https://www.w3.org/TR/vocab-dqv/.

[7] https://www.omg.org/spec/SBVR/1.5/About-SBVR/.
[8] https://www.w3.org/Submission/SWRL/.
[9] https://www.omg.org/spec/PRR/.
[10] https://www.omg.org/dmn/.
[11] https://camunda.com/products/dmn-engine/.
[12] https://www.drools.org/learn/dmn.html.
[13] http://ruleml.org/.

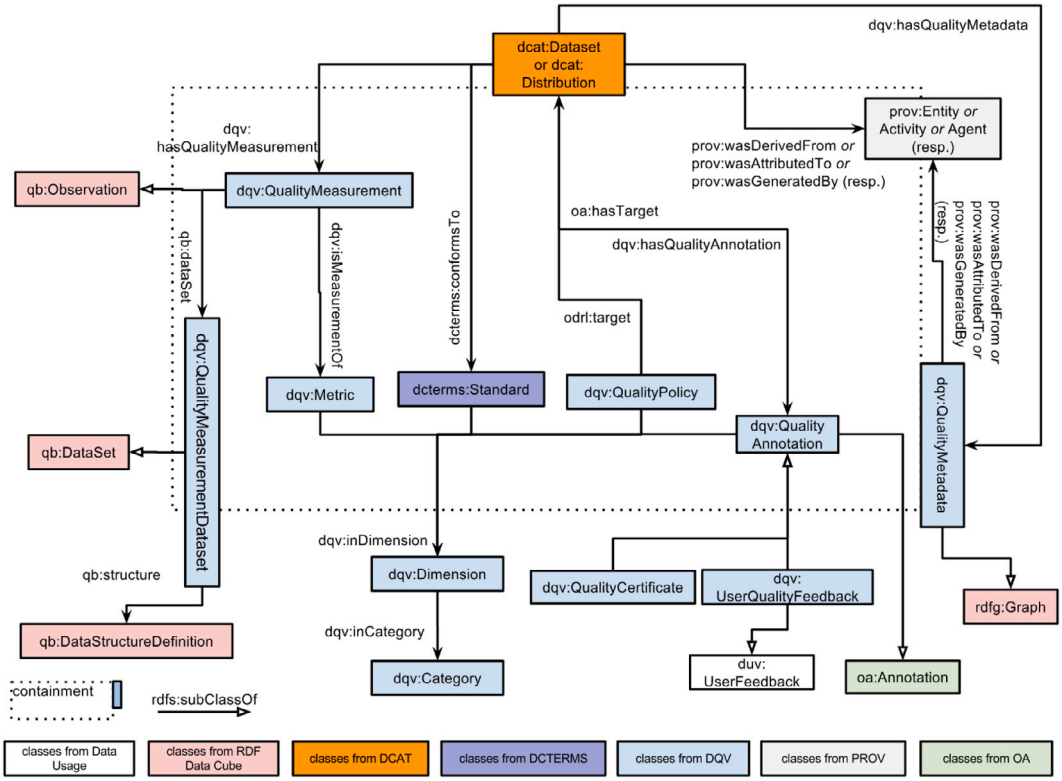


Fig. 2. Data Quality Vocabulary (DQV) of W3C. DQV provides a framework for describing the quality of a dataset. To express these qualities about a dataset, employ five different types of quality information represented by the following classes: *dqv:QualityAnnotation* for describing feedback and quality certificates, *dcterms:Standard* whose main is to represent a standard the dataset or its distribution. *dqv:QualityPolicy* describes a policy governed by data quality concerns. *dqv:QualityMeasurement* whose goal is to provide a metric value about the dataset or distribution that delivers quantitative or qualitative information. And finally, *prov:Entity* describes an entity involved in the provenance of the dataset or distribution. Furthermore, DQV disposes of concepts and properties for data dimension, metrics, category, etc.

- **Semantic definition of data**. Ontology elements support the creation of textual descriptions of the semantics of their classes and properties. Therefore, data providers can identify concepts from other ontologies that can be reused for their own data. For example, in [54], the authors explain how to extract semantic features of images to define their data.
- **Use of semantics data as reference data**. Semantics provides an enormous variety of data from several domains, such as biology, media, and life sciences, which can be used as trusted reference data in data quality monitoring. For example, [50] describes the requirements for legal value and functional dependency in any domain. This work proposes using RDF to format the data. Thus, semantics can be employed to compare data with the trusted reference data.
- **Content integration with ontologies**. Defining hierarchies and class relationships in ontologies ease content analysis at several levels, satisfying their definitions. Besides, the modelisation of equivalence relationships explicitly helps to analyse the data without knowing all synonym relationships. For instance, [55] defined conceptual models in a flexible and extendable way, which allows modelling complex analytical tasks for heterogeneous data sources.

Regarding data quality ontologies, they have been the subject of extensive research in recent years, as data quality has become increasingly important in various domains, such as healthcare [16,56], ground [18], finance [17], and government [57].

Moreover, [58] presents a methodology and an ontology to assess the data quality, but only when the data is in RDF format. Furthermore, other approaches help to represent data quality assessment, such as data Quality Management (DQM) [52,59], Data Cleaning Ontology (DCO) [60], Data Quality Ontology (daQ) [61], Data Quality Vocabulary(DQV) [62] and Fuzzy Quality Data Vocabulary (FQV) [63]. The ontologies above do not help assess the data quality; instead, they publish quality reports in a machine-readable manner. From a different point of view, Reasoning Violation Ontology (RVO) [61] is a dedicated reasoning error ontology that allows processing data issues. [64] evaluates data quality employing semantic rules which can be user-defined in an ontology for data stream applications. In addition, [15] defines concepts and data quality measures in healthcare data. [65] presents an ontology-based data quality framework for relational data stream management systems that include data quality measurement; this work briefly describes data quality concepts through an ontology.

Linked Open Data (LOD) [66] and ontologies are used to describe data quality characteristics due to LOD enables defining the presence of interlinks between datasets and using ontologies as data schemes [67]. LOD allows facing data quality essential issues such as missing data, missing entity relationships, and erroneous data values [68]. Furthermore, LOD can transform data from one format to linked data. However, this transformation can degrade data quality due to various problems, such as errors introduced at the source, parsing values, or interpreting [69]. Data integration in LOD from multiple sources does not continually improve data quality due to contradictory information from the different sources [70]. Regardless of the total number of integrated data sources, quality issues persist at the schema and instance levels [71].

However, many ontologies are inaccessible or are only sparsely described in published scientific works. In addition, the available ontologies do not fulfil our needs completely of creating an ontology for incorporating data quality dimensions, which provides mechanisms for conducting reasoning over the data quality annotations and facilitates the incorporation of the data quality tasks in Big Data workflows.

	BR.DQA		
	Input		Output
F	BR.DQM.Completeness	BR.DQM.Accuracy	BR.DQA
	{Complete, Not complete}	Number	{Suitable, Sufficient, Bad}
1	Complete	100	Suitable
2	-	100	Sufficient
3	-	-	Bad

Fig. 3. Example of the Business Rule in a DMN table used in Data Quality Assessment. Each row is a rule describing the evaluation of the data quality by a set of Business Rule Data Quality Measurement (BR.DQM) combining the results of the measurement of several data quality dimensions.

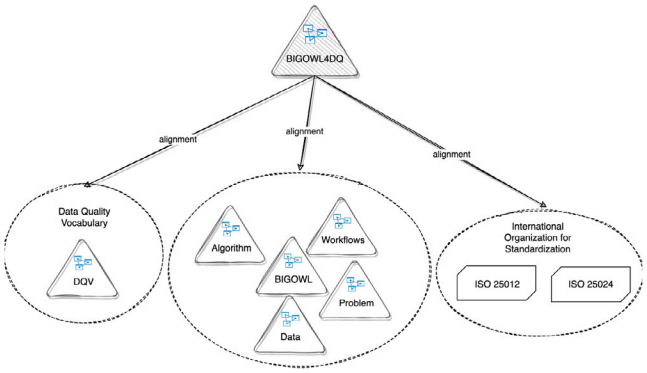


Fig. 4. Overview of the relations of the BIGOWL4DQ ontology. BIGOWL4DQ reuses concepts related to Data Quality and Big Data created previously. Thus, it is aligned with Data Quality Vocabulary, BIGOWL, and with the standards ISO 25012 and 25024.

3. Semantic model of BIGOWL4DQ

One of the main goals of this study is to represent data quality measurement, management concepts, and relationships in Big Data analysis using a formalised language supported by an engine to evaluate data quality rules. For this reason, we opted to extend BIGOWL ontology with the expressivity of DMN4DQ to describe a set of data quality rules in a defined domain and context. To the best of our knowledge, DMN4DQ is the only existing methodology supported by an engine that can assess the data quality of a dataset, based on a set of data quality rules that are machine-readable and processable. This is the key reason why we have based our BIGOWL extension on DMN4DQ.

The proposal of DMN4DQ [25] describes a hierarchical structure which is defined (from bottom to top) as follows: (i) the Business Rule for Data Value (BR.DV) evaluates every data record provided as input of the dataset; (ii) for each data quality characteristic, a Business Rules for Data Quality Measurement (BR.DQM) combines the retrieved outputs of the required BR.DVs as input to generate a data quality measure; (iii) Business Rules for Data Quality Assessment (BR.DQA) uses the outputs of different DMN tables related to the measurement of a dimension (BR.DQM) as input to generate a data quality assessment as shown in the example of Fig. 3; and, (iv) BR.DUD takes the outputs of BR.DQA as input determines the level of usability of each record.

For sharing the data quality rule vocabulary and the data of the Big Data workflow, we propose BIGOWL4DQ. This new ontology is aligned with the W3C Data Quality Vocabulary, which provides an extended group of concepts regarding the quality of the dataset. It also reuses the standards ISO 25012 and 25024 as shown in Fig. 4. Thus, BIGOWL4DQ includes the TBox to represent data quality measurement, management concepts, and relationships. Therefore, these elements can be used later to introduce specific use cases through ABox. Furthermore, our ontology provides a series of Semantic Web Rule Language rules to validate the correctness and completeness of the business rules.

3.1. BIGOWL4DQ: An ontology for data quality management

The main goal of this work is to capture all the semantics needed to define data quality measurement and assessment. For this reason, we opt to design a new OWL 2 ontology extending BIGOWL to describe business rules, data quality characteristics, and quality measures, among others, in the Big Data context. To this end, the standard Ontology 101 development process [28] has been followed, which comprises seven steps:

1. *Determine the domain and scope of the ontology*. The main scope of BIGOWL4DQ is the evaluation and assessment of data quality in big data environments. This scope involves business rules for data values, data quality measurement, assessment, or usability decisions orientated to Big Data.
2. *Consider reusing existing ontologies*. The proposed ontology extends BIGOWL, which has been successfully assessed to define the lifecycle of a workflow, from data reading to the results view. Furthermore, BIGOWL4DQ reuses DQ-Vocab concepts, which provide a metadata model for data quality, e.g. Dimension, Dataset, etc.
3. *Enumerate important terms in the ontology*. Outstanding terms have been selected from the literature on data quality, specifically DQMIM and DQ-Vocab. In addition, terms from the ontologies aligned [62] are incorporated as well as Data Quality Vocabulary. Examples of such terms are *QualityMeasurement* or *DataSet*.
4. *Define the classes and the class hierarchy*. We have followed a top-down approach in developing the class hierarchy. This fact makes it easier to align with BIGOWL and Data Quality Vocabulary, create annotation mappings, and employ a semantic reasoner, among other things. Fig. 5 shows the core classes of the ontology and the hierarchy of the ontology. For example, class *Business Rules* has several subclasses, including *Business Rules Data Usability Decision*, *Business Rules Data Quality Assessment*, *Business Rules Data Quality Measurement*, and *Business Rules Data Value*. BIGOWL4DQ has been developed using Protégé and OWL 2.
5. *Define the properties of classes and slots*. 37 object properties and 24 data properties have been included to relate classes and define attributes. An illustrative set of properties is shown in Table 3, where the class *Clause* is related to the class Condition by means of the property of the object *hasCondition*. Data properties of class *Clause* are *annotation*, or *hasAssociatedHitPolicyCriteria*.
6. *Define the facets of the slots*. This step aims to include cardinality constraints and value restrictions for the ontology's properties. For example, the range of the property *creationDate* is restricted to the date to indicate when the class *BussinesRule* is in its domain.
7. *Create instances*. The instances or individuals in BIGOWL4DQ are specific to the data quality domain. For example, *Completeness* is an instance of the class *DataQualityCharacteristic*. The class *BusinessRule* has a property *hasHitPolicy* (with range *Policy*) to indicate general assertions or guidelines on how a business rule is intended to operate. For example, to validate our proposal, an educational use case from [25] was used to create its rules as instances of BIGOWL4DQ and then assessed them. For example, Fig. 6 shows a partial view of the RDF Graph for the DMN tables for BR.DV.04. The complete set of ontology instances of this case is available at *"Case-study-data-quality-DMN4DQ.owl"* file at the GitHub repository.[14]

Fig. 5 shows the hierarchy to describe the primary data quality ideas. There are defined classes to represent, throughout business rules,

[14] https://github.com/ProyectoAether/BIGOWL4DQ.

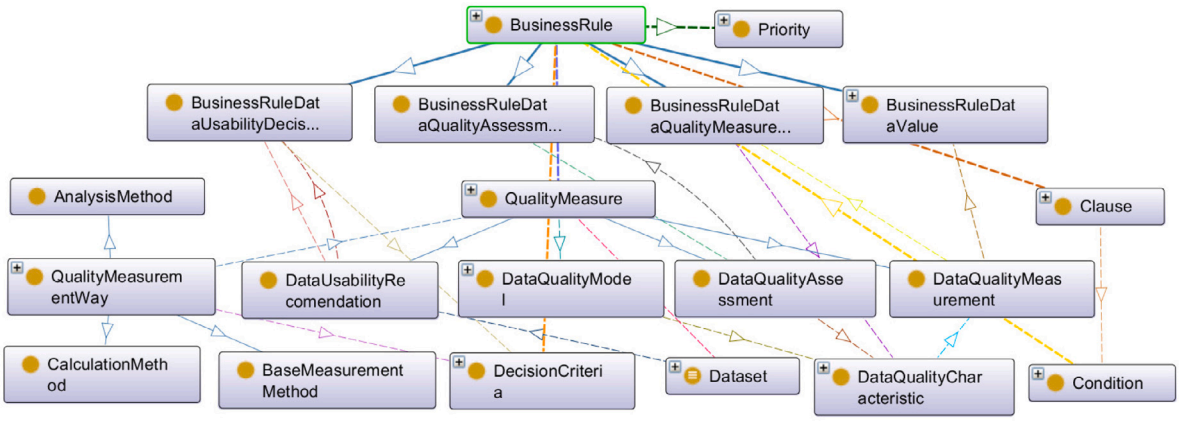


Fig. 5. Overview of the BIGOWL for Data Quality ontology. Subclasses are shown by continuous arrows, whereas dotted arrows denote characteristics. BIGOWL4DQ defines data quality concepts. There are four types of business rules: Data Value, Data Quality Measurement Dimension, Data Quality Dimension, and Data Usability Dimension. Furthermore, it describes Quality Measure concepts, clauses, methods, decision criteria, conditions, etc.

Table 3
Data and object properties to annotate data quality concepts. It can be specified features like business rule domain or clause, hit policy, etc.

Object properties	Description logic	Textual description
domain	∃ domain Thing ⊑ BusinessOutput⊤ ⊑ ∀ domain (PrimitiveType ⊔ BusinessOutput)	Specifies the domain of the business output, ensuring it belongs to the PrimitiveType or BusinessOutput category.
generates	∃ generates Thing ⊑ QualityMeasurementWay⊤ ⊑ ∀ generates QualityMeasure	Represents the method or way by which quality measurements are generated, falling under the QualityMeasurementWay or QualityMeasure categories.
hasAssociated	∃ hasAssociated Thing ⊑ Dataset⊤ ⊑ ∀ hasAssociated DataUsabilityRecomendation	Links the property to its associated dataset, falling under the Dataset or DataUsabilityRecommendation category.
hasClause	∃ hasClause Thing ⊑ BusinessRule⊤ ⊑ ∀ hasClause Clause	Establishes a connection between a business rule and its associated clause. It ensures that clauses belong to the BusinessRule or Clause category.
hasCondition	∃ hasCondition Thing ⊑ Clause⊤ ⊑ ∀ hasCondition Condition	Indicates the conditions associated with a clause, ensuring they belong to the Clause or Condition category.
hasCriteriaDecision	∃ hasCriteriaDecision Thing ⊑ BusinessRuleDataUsabilityDecision⊤ ⊑ ∀ hasCriteriaDecision DecisionCriteria	Connects a business rule to its data usability decision criteria, ensuring they fall under the BusinessRuleDataUsabilityDecision or DecisionCriteria category.
involves	∃ involves Thing ⊑ QualityMeasurementWay⊤ ⊑ ∀ involves DecisionCriteria	Links a property to the decision criteria involved in quality measurement, ensuring they belong to the QualityMeasurementWay or DecisionCriteria category.
Data properties	**Description logic**	**Textual description**
annotation	∃ annotation Datatype ⊑ Clause⊤ ⊑ ∀ annotation	Describes additional information or metadata associated with a clause, falling under the Datatype or Clause category.
businessRulesSource	∃ businessRulesSource ⊑ BusinessRule⊤ ⊑ ∀ businessRulesSource	Specifies the source of business rules, ensuring they belong to the BusinessRule category.
creationDate	∃ creationDate ⊑ BusinessRule ⊤ ⊑ ∀ creationDate Datatype	Represents the date when a business rule was created, falling under the BusinessRule or Datatype category.
hasAssociatedHitPolicyCriteria	∃ hasAssociatedHitPolicyCriteria ⊑ Clause⊤ ⊑ ∀ hasAssociatedHitPolicyCriteria Datatype	Links a property to its associated hit policy criteria, falling under the Clause or Datatype category.
name	∃ name ⊑ Field	Represents the name associated with a field.
isCorrect	∃ isCorrect⊑ BusinessRule⊤ ⊑ ∀ isCorrect Datatype	Indicates whether a business rule is correct, falling under the BusinessRule or Datatype category.
isCompleteField	∃ isCompleteField⊑ Field⊤ ⊑ ∀ isCompleteField Datatype	Specifies whether a field is complete, falling under the Field or Datatype category.
isCompleteCondition	∃ isCompleteCondition⊑ Condition⊤ ⊑ ∀ isCompleteCondition Datatype	Indicates whether a condition is complete, falling under the Condition or Datatype category.
isCompleteClause	∃ isCompleteClause⊑ Clause⊤ ⊑ ∀ isCompleteClause Datatype	Specifies whether a clause is complete, falling under the Clause or Datatype category.
isComplete	∃ isComplete⊑ BusinessRule⊤ ⊑ ∀ isComplete Datatype	Indicates whether a business rule is complete, falling under the BusinessRule or Datatype category.

the user requirements for the data in a context of use to generate a recommendation on the usability of the data. There are four types of business rules: Data Value, Data Quality Measurement Dimension, Data Quality Dimension, and Data Usability Dimension. To measure the quality of the different business rules, the class Data Quality Characteristic, measured by the class Quality Measure, uses different ways annotated with the class Quality Measurement Way and methods determined by the class Calculation Method. Furthermore, more classes

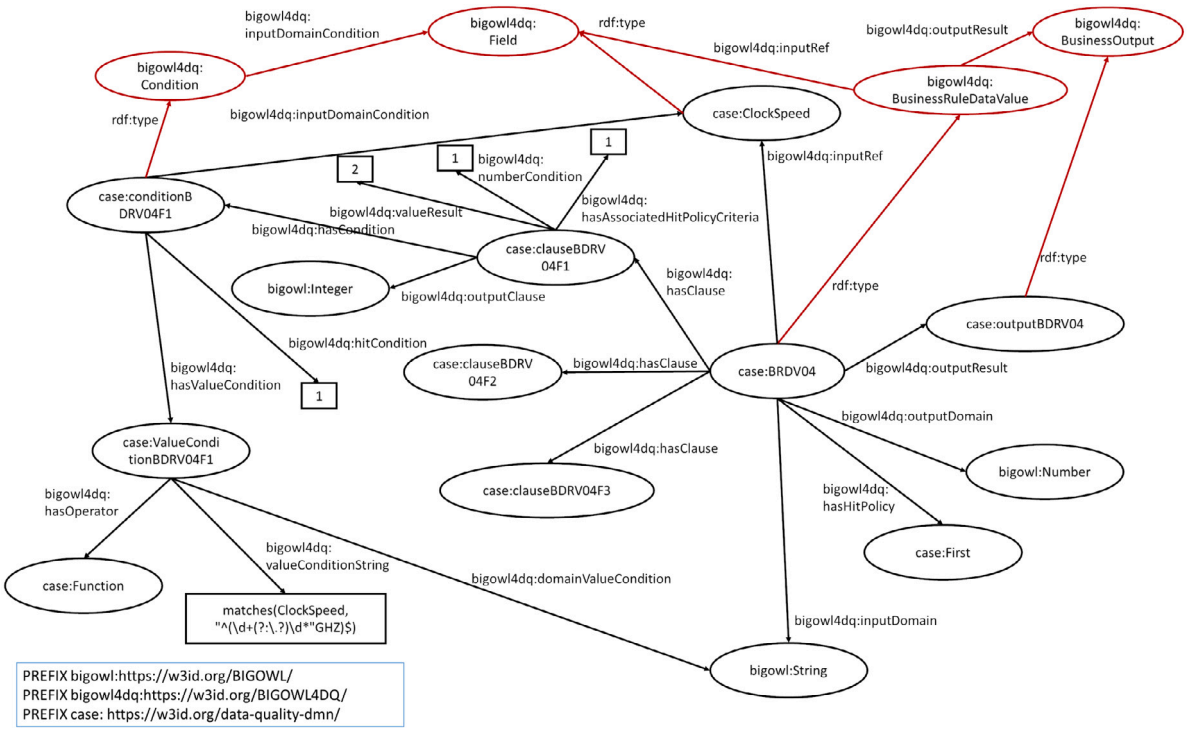


Fig. 6. General overview of an example of BIGOWL4DQ instances for one DMN table. In red, the ellipses depict the classes described in our ontology. In addition, in black ellipses, the samples created to illustrate the example of the DMN table are shown. The arrows represent the properties of the data and objects.

are indicated to describe other concepts like clauses, decision criteria, conditions, etc. The complete ontology is developed in *"bigowl4dq.owl"* file and available in the GitHub repository[14].

3.2. Reasoning rule framework

The high number of business rules for data quality can imply a complex validation process. Some of the most common validations are based on analysing correctness and completeness. In the context of DMN tables, Dumas et al. [72] proposed a formal validation that analyses possible missing and overlapping rules, duplicate rules, conflicting rules, shadowed rules, types of expression, correct use of enumerations, and correctly connected requirement graphs. The proposal is supported by two tools to validate the DMN tables: dmn-js[15] and dmn-check.[16] In contrast with those tools, our proposal validates DMN tables and any business rules.

This section describes a set of SWRL rules to control correctness and completeness using a semantic-based approach. The SWRL rules can consider the data quality rule aspects and the semantic context of the problem under study. BIGOWL4DQ defines a set of SWRL rules on top of the OWL 2 ontology to derive new information from existing knowledge. These rules are used by semantic reasoning tasks primarily concerned with verifying the validity of the rules, such as identifying rules that are (in)compatible with the input domain, clauses, conditions, etc. Another possible use is to check the completeness of the rules, i.e., every potential input configuration trigger has at least one associated rule. Thus, the primary purpose of these rules is to create well-formed business rules. These SWRL are included in the complete ontology (*"bigowl4dq.owl"*) available in the GitHub repository[14].

The following is a list of these rules to check the validity of several concerns of the identified business rules for data quality evaluation and assessment:

- SWRL Rule for checking the correctness in Business Rules Data Value. This rule is used to check the compatibility between the domains of the business rules and the field in which the data are applied. In addition, it is assessed that the rules contain clauses and define their conditions and outputs.

Correctness of Business Rule Data Value

```
bigowl4dq:BusinessRuleDataValue(?br) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:outputClause(?cl, ?outCl) ^
bigowl4dq:inputRef(?br, ?ref) ^
bigowl4dq:hasFieldDomain(?ref, ?indo) ^
bigowl4dq:outputDomain(?br, ?od) ^
bigowl4dq:inputDomain(?br, ?indo) ^
bigowl4dq:outputResult(?br, ?ores) ^
bigowl4dq:domain(?ores, ?od) ->
bigowl4dq:isCorrect(?br, true)
```

- SWRL Rule for checking the correctness in Business Rules Data Quality Measurement Dimension. This rule examines the harmony between the input data domain of the Data Quality Measurement Dimension (DQMD) rules and the output data domain of the Data Value rules. Both domains must match because the output of the data quality rules for data values (BR.DV) is the input of the DQMD rules. In addition, it checks that the clauses, conditions, and outcomes are well-defined in the DQMD rules.

[15] dmn-js: http://dmn.cs.ut.ee.
[16] dmn-check: https://github.com/red6/dmn-check#validations.

Correctness in Business Rules Data Quality Measurement

```
bigowl4dq:BusinessRuleDataValue(?br) ^
bigowl4dq:outputResult(?brDV, ?or) ^
bigowl4dq:outputDomain(?br, ?od) ^
bigowl4dq:inputDomain(?br, ?or) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:outputClause(?cl, ?outCl) ->
bigowl4dq:isCorrect(?br, true)
```

- SWRL Rule for checking the correctness in Business Rules Data Quality Assessment. This rule checks for compatibility between the input data domain of the Data Quality Assessment (DQA) rules and the DQMD's output data domain. The result of the DQMD rules is the input of the DQA rules. Therefore, both domains must be compatible. It also ensures that the clauses, conditions, and outcomes of the DQA rules are well-defined.

Correctness in Business Rules Data Quality Assessment

```
bigowl4dq:BusinessRuleDataQualityAssessment(?br) ^
bigowl4dq:outputResult(?br, ?or) ^
bigowl4dq:BusinessRuleDataUsabilityDecision(br) ^
bigowl4dq:outputDomain(?br, ?od) ^
bigowl4dq:inputDomain(?br, ?or) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:outputClause(?cl, ?outCl) ->
bigowl4dq:isCorrect(?br, true)
```

- SWRL Rule for checking the correctness in Business Rules Data Usability Decision. This rule verifies that the input data domain of the Data Usability Decision (DUD) rules and the output data domain of the DQA are compatible. Both domains must be identical because the output of the DQA rules equals the input of the DUD rules. It also adequately defines the DUD rules' clauses, conditions, and results.

Correctness in Business Rules Data Usability Decision

```
bigowl4dq:BusinessRuleDataUsabilityDecision(?br) ^
bigowl4dq:outputResult(?br, ?or) ^
bigowl4dq:outputDomain(?br, ?od) ^
bigowl4dq:inputDomain(?br, ?or) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:outputClause(?cl, ?outCl) ->
bigowl4dq:isCorrect(?br, true)
```

- Rules for checking the completeness in a condition. This family of rules proves that the range of values defined in a clause condition covers any input value. There are rules for checking the interval of numeric, Boolean, or String values.

For instance, the following rule verifies that a maximum value of an interval of a numeric field is covered; it is checked that the maximum value defined in the interval of a condition is equal to the value specified in the other conditions. This means that the first condition defines its interval as $[1, 3]$ and the second condition as >3. As a consequence, all values greater than or equal to 1 are covered by those conditions. Rules have been described for all different combinations.

Case 1: Completeness in a condition

```
bigowl4dq:BusinessRuleDataValue(?br) ^
bigowl4dq:inputRef(?br, ?inRef) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:inputDomainCondition(?cond, ?inRef) ^
bigowl4dq:hasValueCondition(?cond, ?valCond) ^
bigowl4dq:inputDomainCondition(?cond2, ?inRef) ^
bigowl4dq:hasValueCondition(?cond2, ?valCond2) ^
bigowl4dq:hasOperator(?valCond,
bigowl4dq:BetweenValues) ^
bigowl4dq:hasOperator(?valCond2,
bigowl4dq:EqualOrGreater) ^
bigowl4dq:rangeMax(?valCond, ?max) ^
bigowl4dq:valueConditionNumber(?valCond2, ?max) ->
bigowl4dq:isCompleteCondition(?cond, true)
```

In the case where the condition contains the default value (e.g., "-"), this means that the condition will meet any input value; therefore, this condition will be triggered by any input value; then, the condition is established as complete.

Case 2: Completeness in a condition

```
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:hasAssociatedHitPolicyCriteria(?cl, ?hit) ^
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:hitCondition(?cond, ?hit) ^
bigowl4dq:hasValueCondition(?cond, ?valCond) ^
bigowl4dq:valueConditionString(?valCond, ?str) ^
swrlb:equal(?str, "-") ->
bigowl4dq:isCompleteCondition(?cond, true)
```

- Rules for checking the completeness in a clause. One or more conditions form a clause. Two rules have been defined to assess the completeness of a clause.

The first case is when a clause contains only one condition. In this case, the rule checks the completeness of the condition and inherits its value, as shown below.

Case 1: Completeness in a clause

```
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:isCompleteCondition(?cond, ?valCond) ^
swrlb:equal(?valCond, true) ^
bigowl4dq:numberCondition(?cl, ?num) ^
swrlb:equal(?num, 1) ^
bigowl4dq:hitCondition(?cond, ?hit) ^
bigowl4dq:hasAssociatedHitPolicyCriteria(?cl, ?hit)
-> bigowl4dq:isCompleteClause(?cl, true)
```

The second case is when a clause has two or more conditions; in this case, the rule assesses the completeness of each pair of conditions, and only when all its conditions meet the completeness, the clause does it.

Case 2: Completeness in a clause

```
bigowl4dq:hasCondition(?cl, ?cond) ^
bigowl4dq:hasCondition(?cl, ?cond2) ^
bigowl4dq:numberCondition(?cl, ?num) ^
swrlb:greaterThan(?num, 1) ^
differentFrom(?cond, ?cond2) ^
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:isCompleteCondition(?cond, ?valCond) ^
swrlb:equal(?valCond, true) ^
bigowl4dq:isCompleteCondition(?cond2, ?valCon2) ^
swrlb:equal(?valCond2, true) ^
bigowl4dq:hitCondition(?cond, ?hit) ^
bigowl4dq:hasAssociatedHitPolicyCriteria(?cl, ?hit) ^
bigowl4dq:hitCondition(?cond2, ?hit) ->
bigowl4dq:isCompleteClause(?cl, true)
```

- Rule for checking the completeness in a business rule. A business rule is triggered with any input value only if it contains at least one completeness clause. Thus, if we find at least one completeness clause, then the completeness of the business rules is ensured.

Completeness in a business rule

```
bigowl4dq:hasClause(?br, ?cl) ^
bigowl4dq:isCompleteClause(?cl, true) ->
bigowl4dq:isComplete(?br, true)
```

3.2.1. SWRL validation

Two experiments have been conducted from the educational use case presented in [25] have been conducted to validate the proposed semantic approach. In both tests, the data quality rules are described as DMN business rules. The experiments assess the correctness and completeness of the DMN rules through the SWRL rules described in BIGOWL4DQ. The use case is based on a catalogue of servers for private clouds. The data provide information about the different features of the servers, such as the amount of RAM, storage capacity, clock speed, etc. There are defined SWRL rules to evaluate the correctness of the business rules. The DMN tables are used to validate the SWRL rules in this case. Fig. 7(a) shows how the SWRL rule is used to check the correctness of a data quality rule for the value of the data. In the blue area, the SWRL rule checks to see if the business rule comprises at least a clause, a condition, and an input field for the condition. The orange portion of the SWRL rule also verifies that the inputs and outputs of the field, clause, condition, and business rules are consistent. The instances generated for this table produce as a result the assertion of a "true" value for the data property *isCorrect* for the instance of a *BusinessRuleDataValue* class (*case:BRDV04* in Fig. 6). After the evaluation of all the correctness rules, those rules that do not include the "true" value for the *isCorrect* property are considered incorrect, while there are no new assertions.

In the case of the completeness of a rule, it is split into different levels. First, it is checked that the conditions in the various clauses cover all the domains; then, it is checked that at least one clause of the business rule is complete; if so, the rule fulfils the completeness. Fig. 7(b) shows an example of the use of a SWRL rule to assess completeness at the condition level. Thus, the first section of the SWRL rule, shown in blue, looks to see if the business rule has a hit policy and at least one condition. Additionally, the orange portion of the SWRL rule verifies that the condition includes at least a default value, in this example, "-". After the evaluation of all the completeness rules, those rules that do not include the "true" value for the *isCompleteClause* property are considered incomplete; meanwhile, there are no new assertions.

4. Use case

Sustainability assessment is a significant concern in smart farming [73]. As observed in Fig. 8, it is possible to take advantage of a Big Data pipeline [74,75] to optimise specific smart farm decision-making processes for sustainability, e.g., to optimise irrigation in specific periods with low humidity levels using Artificial Intelligence (AI) techniques. This decision-making process conveys the production of several indicators [76] using AI techniques to transform some raw data in agricultural businesses [77]. However, sustainability assessment presents some multi-criteria problems when producing these indicators. Data quality has been shown to be one of these criteria for selecting the corresponding arrows to enable farmers to make sound decisions [78]. Many data quality problems can affect the reliability of these indicators; for example, incomplete or missing values can lead to undesirable results, or data outside the allowed values may lead to inaccurate results [79]. We confirm that the enumeration of the latter problems in terms of *incomplete or missing values* could be too ambiguous, since it is not stated how many incomplete data can be accepted to process the data reliably.

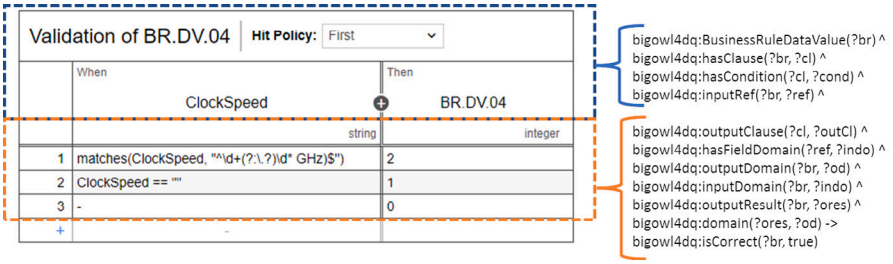
Furthermore, merely enumerating the list cannot help identify how to better design more directed corrective actions. To achieve this aim, identifying the data quality characteristics that represent the data quality requirements for each use case can help better support the data preparation phases. These data quality problems are usually rooted in difficulties in understanding and acting on the data and in adjusting the data before applying any technique [80]. Therefore, it is essential to provide users with practical and easy-to-understand information on the level of data quality. In most contexts, this information can be provided in terms of usability [25,81] before starting any decision-making process to prevent a data quality problem that leads to a misleading decision [82].

To illustrate the impact of data quality, we introduce the case study presented in Fig. 8. This case is based on a real dataset for a smart farm provided in [83–85]. This dataset collected information about a smart farm from 42 locations where several sensors send information (with different frequencies). Sensors measure hourly and daily volumetric water content, soil temperature, and bulk electrical conductivity at depths of 30, 60, 90, 120, and 150 cm across the farm. These data are stored in plain text (raw data) separated by sensors, day, and hour. The complete dataset (available at[17]) consisted of 1,048,581 records. These data records gather information for the following features:

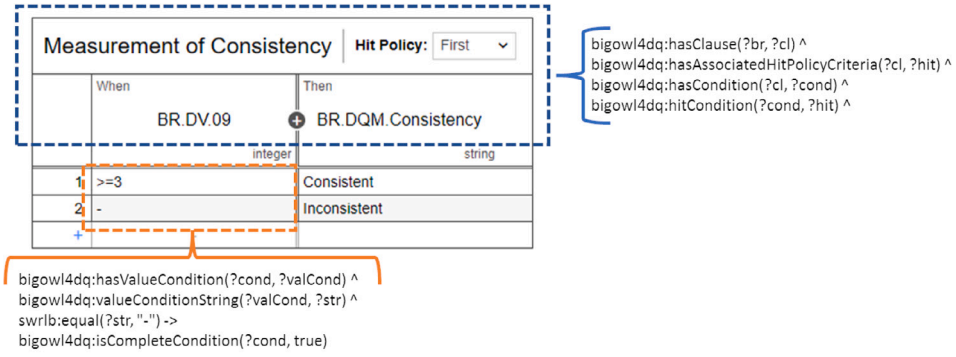
- *Location* is the name of the sensor.
- *Date* of data reading.
- *Time* of data reading.
- VW_30 cm is the humidity at 30 cm depth.
- VW_60 cm is the humidity at 60 cm depth.
- VW_90 cm is the humidity at 90 cm depth.
- VW_120 cm is the humidity at 120 cm depth.
- VW_150 cm is the humidity at 150 cm depth.
- T_30 cm is the temperature at a depth of 30 cm.
- T_60 cm is the temperature at a depth of 60 cm.
- T_90 cm is the temperature at a depth of 90 cm.
- T_120 cm is the temperature at a depth of 120 cm.
- T_150 cm is the temperature at a depth of 150 cm.

These data can be used to automate sensor re-calibration [86]. A significant concern in this scenario is the data acquisition in which several heterogeneous sensors are distributed across a location, and precise and accurate measurements are desirable. Due to the sensors' heterogeneity, their sensors' calibration can vary, creating problems in readings, and therefore in the quality levels of the data obtained. Of course, it can affect further processing based on the data on a non-well-calibrated sensor.

[17] https://doi.org/10.15482/USDA.ADC/1349683.



(a) Overview SWRL rule for checking the correctness of the Business Rule Data Value DMN table. Each row in the table represents a rule, in this case, relative to the *Clock Speed* variable. The first column describes the rules themselves, and the second column describes their results when evaluated. On the right is shown the SWRL rule that assesses the correctness of the Business Rule Data Value.



(b) Overview SWRL rule for assessing the completeness of the Business Rule Data Value DMN table. Each row in the table represents a rule. The first column represents the value from evaluating a Business Rule Data Value. Furthermore, the second one describes the consistency. On the right is shown the SWRL rule that evaluates the completeness of the Business Rule Data Value.

Fig. 7. SWRL rules to check properties of a business rule described by a DMN table. At the top (a), the SWRL rule assesses the correctness of a Business Rule Data Value. The bottom (b) represents how to evaluate a rule's completeness.

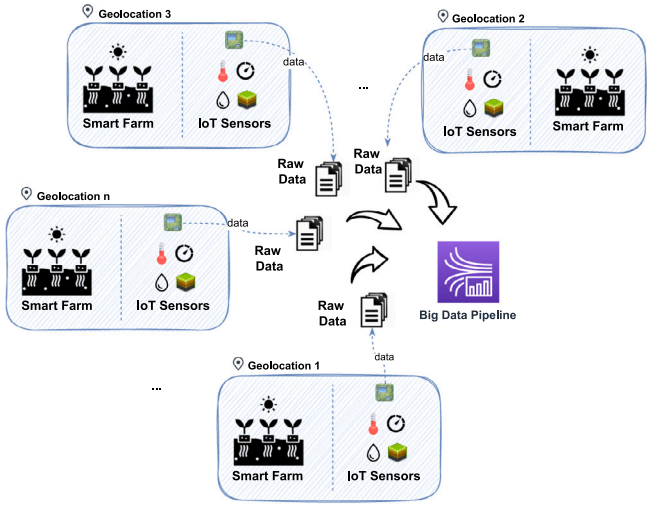


Fig. 8. Smart Farm scenario where data is collected from different locations. Each one has a group of sensors that provides raw data like temperature, humidity, etc. Finally, all data are analysed using a Big Data pipeline.

After inspecting the data, the dataset included some quality-related problems [83–85]. From a classical point of view of data preparation, the surrounding actions are directed at cleaning the data. This data cleaning will probably include validating if known data errors are required to be analysed for the context of use rather than previously determining if data have an adequate level of quality enough for the task at hand. In this case, this would involve first determining the risk appetite of the data scientists in charge of designing the sensor re-calibration model. The risk appetite will be used to delimit whether or not a data record should be used for the task at hand. In DMN4DQ, risk appetite is represented by means of the corresponding business rules that define the data quality requirement of users.

Evaluating the potential impact of inadequate levels of quality of the data used (i.e., *Time*, *Date*, *Location*, VW_30 cm, VW_60 cm, etc.) and based on the recalibration process [86], we provide some suggestions on the levels of quality of these features required for optimal use for recalibration according to experts. In this sense, we identified two data quality dimensions (*completeness* and *accuracy*) as criteria to better examine the data pattern to recalibrate the sensors. Taking advantage of the application of the BR4DQ methodology [42], the set of business rules for the data values (see Table 4 introduced in [85]) was classified for each attribute of the data and grouped for each characteristic of the quality of the data considered during the assessment (see Table 5).

Table 6 gathers the various business rules to measure the characteristics of data quality selected for the case study.

Table 7 gathers the corresponding business rules for the assessment of the quality of the data from the case study.

Table 8 shows the business rules –representing the risk appetite of the organisation– to determine whether the data should be used or not for the task at hand of recalibrating sensors. It shows the results of applying the BR.DUD to the dataset and gathers the results of the determination of the usability of the data.

Fig. 9 shows the instance of the BR.DV09 table using BIGOWL4DQ. This instance links to the inputs, output domains, and the four clauses. This instance is an example of a correct and complete rule, so the reasoner will end up adding the facts *isCorrect* and *isComplete* (applying the SWRL rules).

In this use case, SWRL validation has been performed; for instance, Fig. 10 depicts a DMN table of assessment that fails to fulfil the rule of

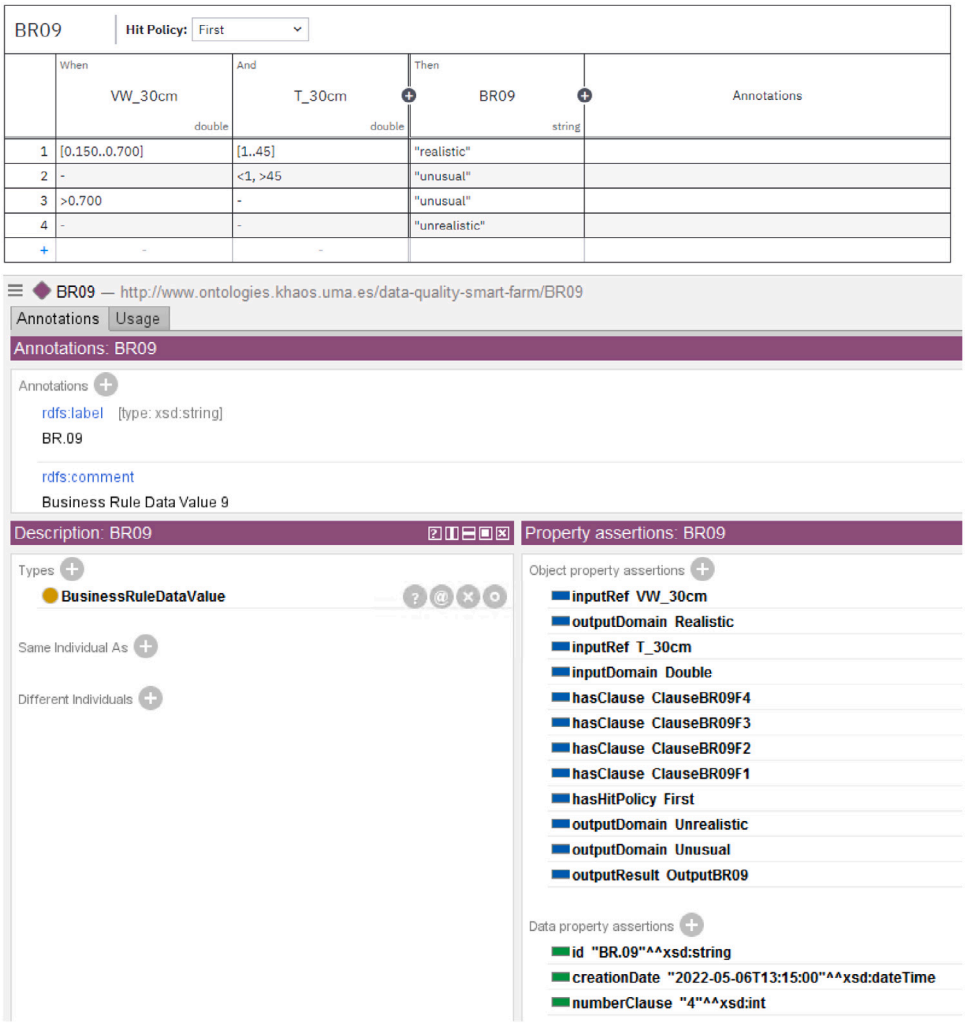


Fig. 9. At the top is depicted the DMN table of the business rule data value BR.DV09. Its annotation at the bottom includes data and object properties with BIGOWL4DQ using the tool Protégé.

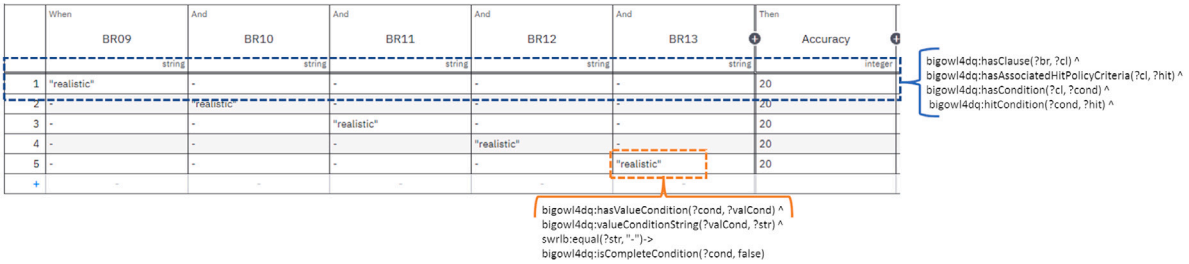


Fig. 10. DMN table (BR.DQM for Accuracy) fails to fulfil the rule of completeness. The blue colour is used to show the part of the SWRL rule that assesses the completeness and passes it. Additionally, the orange colour describes the part of the rule that evaluates the completeness and does not meet it.

completeness; the rule detects that there are no rows with all its cells with the default value "-". In addition, the range of values in the table is not completely covered.

All annotations for this use case using the BIGOWL4DQ ontology together with the SWRL rules are available in the GitHub repository[14] in the file *"Case-study-data-quality-smart-farm.owl"*. Furthermore, the DMN tables can be found in the file *"Case-study-dmn-smart-farm.dmn"*.

5. Threats to validity

Evaluation of threats to validity is critical to ensure the quality of the study, and evaluation of threats to validity is critical. Following the guidance published in [87], four aspects of validity should be considered:

- *Construct validity:* This aspect concerns the degree to which the application of constructs is justified about research objectives and questions [88]. The main goal is to create an ontology that improves reasoning capabilities concerning data quality dimensions for Big Data workflows. For this purpose, we have developed the BIGOWL4DQ ontology. The rationale for constructing BIGOWL4DQ is based on two previous works, the DMN4QD methodology [25] and the BIGOWL [2] ontology. Furthermore, we have used the DMN standard to describe business rules for

Table 4
Description of the business rule data value per data attribute. The first column indicates the data features. The second one indicates the rule where the feature has been employed. And finally, the third column is a description of the business rule.

Features	Business rule ID	Business rule statement
Location	BR.DV1	Location contains data other than null, empty or blank.
Date	BR.DV2	Date contains data other than null, empty, or blank.
Time	BR.DV3	Time contains data other than null, empty, or blank.
VW_30 cm, T_30 cm	BR.DV4	Volumetric and Temperature sensors at 30 cm contain data other than null or "NA".
VW_60 cm, T_60 cm	BR.DV5	Volumetric and Temperature sensors at 60 cm contain data other than null or "NA".
VW_90 cm, T_90 cm	BR.DV6	Volumetric and Temperature sensors at 90 cm contain data other than null or "NA".
VW_120 cm, T_120 cm	BR.DV7	Volumetric and Temperature sensors at 120 cm contain data other than null or "NA".
VW_150 cm, T_150 cm	BR.DV8	Volumetric and Temperature sensors at 150 cm contain data other than null or "NA".
VW_30 cm, T_30 cm	BR.DV9	Volumetric and Temperature sensors at 30 cm are in different ranges: (1) if the humidity is in the range of 0.150 to 0.700 and the temperature between 1 and 45 degrees; (2) in the case where the temperature is below 1 degree or above 45; and, (3) in the case where the humidity is above 0.700.
VW_60 cm, T_60 cm	BR.DV10	Volumetric and Temperature sensors at 30 cm are in different ranges: (1) it is "realistic" when the humidity is in the range of 0.150 to 0.700 and the temperature between 1 and 45 degrees; (2) it is "unusual" when the temperature is below 1 degree or above 45, also when the humidity is above 0.700; otherwise, (3) it is "unrealistic".
VW_90 cm, T_90 cm	BR.DV11	Volumetric and Temperature sensors at 30 cm are in different ranges: (1) it is "realistic" when the humidity is in the range of 0.150 to 0.700 and the temperature between 1 and 45 degrees; (2) it is "unusual" when the temperature is below 1 degree or above 45, also when the humidity is above 0.700; otherwise, (3) it is "unrealistic".
VW_120 cm, T_120 cm	BR.DV12	Volumetric and Temperature sensors at 30 cm are in different ranges: (1) it is "realistic" when the humidity is in the range of 0.150 to 0.700 and the temperature between 1 and 45 degrees; (2) it is "unusual" when the temperature is below 1 degree or above 45, also when the humidity is above 0.700; otherwise, (3) it is "unrealistic".
VW_150 cm, T_150 cm	BR.DV13	Volumetric and Temperature sensors at 30 cm are in different ranges: (1) it is "realistic" when the humidity is in the range of 0.150 to 0.700 and the temperature between 1 and 45 degrees; (2) it is "unusual" when the temperature is below 1 degree or above 45, also when the humidity is above 0.700; otherwise, (3) it is "unrealistic".

data quality and the Data Quality Vocabulary (DQV) by W3C and the two standards ISO/IEC 25012 and 25024 for the extension of data quality dimensions. Although these standards give BIGOWL4DQ very solid baselines, it has been impossible to study

Table 5
Business rules for data value and data quality characteristics. The first column indicates the data quality characteristics. The second one shows the list of rules employed by the Business Rules Data Value. And finally, the third column is a description of the data quality characteristics.

DQ characteristics	Business rule for data value	Description
Completeness	BR.DV1, BR.DV2, BR.DV3, BR.DV4, BR.DV5, BR.DV6, BR.DV7, BR.DV8	Detect missing relevant data from the dataset that may lead to undesirable results.
Accuracy	BR.DV9, BR.DV10, BR.DV11, BR.DV12, BR.DV13	Detect the values collected by the sensors are not reliable due to the extreme values.

Table 6
Description of business rules for data quality measurement for completeness and accuracy characteristics.

DQ characteristics	Business rule ID	Business rule statement
Completeness	BR.DQM.1	Location, data, time, and volumetric (VW_30 cm) and temperature (T_30 cm) sensors contain values. (BR.DV1,BR.DV2,BR.DV3,BR.DV4)
Completeness	BR.DQM.2	Location, data, time, and volumetric (VW_60 cm) and temperature (T_60 cm) sensors contain values. (BR.DV1,BR.DV2,BR.DV3,BR.DV5)
Completeness	BR.DQM.3	Location, data, time, and volumetric (VW_90 cm) and temperature (T_90 cm) sensors contain values. (BR.DV1,BR.DV2,BR.DV3,BR.DV6)
Completeness	BR.DQM.4	Location, data, time, and volumetric (VW_120 cm) and temperature (T_120 cm) sensors contain values. (BR.DV1,BR.DV2,BR.DV3,BR.DV7)
Completeness	BR.DQM.5	Location, data, time, and volumetric (VW_150 cm) and temperature (T_150 cm) sensors contain values. (BR.DV1,BR.DV2,BR.DV3,BR.DV8)
Accuracy	BR.DQM.1	Volumetric (VW_30 cm) and temperature (T_30 cm) sensors provided "realistic" values (BR.DV9).
Accuracy	BR.DQM.2	Volumetric (VW_60 cm) and temperature (T_60 cm) sensors provided "realistic" values (BR.DV10).
Accuracy	BR.DQM.3	Volumetric (VW_90 cm) and temperature (T_90 cm) sensors provided "realistic" values (BR.DV11).
Accuracy	BR.DQM.4	Volumetric (VW_120 cm) and temperature (T_120 cm) sensors provided "realistic" values (BR.DV12).
Accuracy	BR.DQM.5	Volumetric (VW_150 cm) and temperature (T_150 cm) sensors provided "realistic" values (BR.DV13).

Table 7
Statement description of how each business data quality assessment is evaluated.

Business rule ID	Business rule statement
BR.DQA.1	It is "suitable" for those records that have a Completeness measurement greater than or equal to 5 and accuracy greater than or equal to 100. All the sensors provided complete and accurate data.
BR.DQA.2	It is "enough quality" when there is at least a value equal to or greater than 3 in the Completeness measurement and greater than or equal to 60 in the Accuracy measurement. Thus, three or four sensors provided complete and accurate readings.
BR.DQA.3	It is "bad quality" when complete readings from one or two sensors but with a precision below 60.
BR.DQA.4	It is "non-useable" in any other case.

Table 8
Description of business rule data decision to determine the level of usability.

Business rule ID	Business rule statement	Recommendation
BR.DUD.1	Data recorded assessed as "suitable" or "enough quality"	"Use"
BR.DUD.2	Otherwise	"Do not use"

the entire state of the art; hence there may be approaches in the literature that can be used to complement BIGOWL4DQ.

- *Internal validity:* This aspect is related to the quality of the study, as this aspect is highly dependent on the study procedures and the strictness of their execution, so this aspect depends on the efficiency of the study. To avoid possible bias in the construction of the ontology, we have followed the standard Ontology development process 101 [28]. The use of this standard provides a piece of trustworthiness in the development process.
- *External validity:* This aspect is related to the possibility of generalising the results and is of interest to others outside of the study. BIGOWL4DQ is prepared to be used in any Big Data workflow in which we need to measure and assess the data quality's usability level. We have developed BIGOWL4DQ using as a reference the DMN4DQ results, ensuring that it is applicable in an educational case study. Then, BIGOWL4DQ has also been validated in the context of a practical use case of smart farms related to soil sensor networks. Although the application for two points is not a proof of generalisation, we provided all the resources to enable the replication of the case studies presented in the paper. To facilitate generalisation, we offer the BIGOWL4DQ ontology, the implementation of the SWRL rules, and the instances of the two case (educational and realistic) studies for the community in the available repository. This enables anyone to use BIGOWL4DQ and apply it to the context.
- *Conclusion validity:* This aspect aims to reach relevant conclusions through a rigorous and repeatable treatment. As mentioned above, we provided access to all the material to enable replication, and also the BIGOWL4DQ ontology is provided. With this, we achieve a twofold purpose, the reproduction of case studies, and we open our approach to the community for the application in any context.

6. Discussion and conclusions

Considering data quality in the Big Data workflow, we propose BIGOWL4DQ, an ontology to describe business data quality rules, characteristics, measurements, and assessments. This proposal not only facilitates the integration of any data set in a Big Data analytics process for a later application of AI algorithms, but also provides the required mechanisms for reasoning in the set of business quality rules. Our proposal represents actionable knowledge to automate the process of better-supporting data preparation and to reduce the complexity of defining the main concepts of Data Quality management when designing Artificial Intelligence workflows and integrated reasoning capacities. Furthermore, BIGOWL4DQ has been validated in two case studies, including a smart farming case study within an Artificial Intelligence analysis and an academic use case.

Using reasoning that takes advantage of axioms and SWRL rules enables the discovery of implicit knowledge hidden in the expert knowledge expressed in the ontology. The designed process also allows for the identification of correct and complete business rules. Thus, any rule proved to be correct and complete is extended with explicit data properties showing this fact. Under the Open-World Assumption, any rule not annotated as correct or complete cannot be directly classified as incorrect or incomplete. However, from a practical point of view, we will consider that these rules could be incorrect or incomplete; meanwhile, experts do not add new knowledge to the ontology.

In future work, we plan to develop software tools to automatically create BIGOWL4DQ instances from data quality business rules. Thus, the use of any tool creating data quality business rules for Data Quality will be possible to automatically translate to OWL 2. This translation based on BIGOWL will also enable the extension of TITAN [89] to consider the Data Quality dimensions in the design of Big Data workflows.

CRediT authorship contribution statement

Cristóbal Barba-González: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Ismael Caballero:** Conceptualization, Data curation, Investigation, Writing – original draft, Writing – review & editing, Software, Formal analysis. **Ángel Jesús Varela-Vaca:** Conceptualization, Validation, Visualization, Writing – original draft, Writing – review & editing, Investigation, Methodology, Data curation, Formal analysis. **José A. Cruz-Lemus:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **María Teresa Gómez-López:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Investigation. **Ismael Navas-Delgado:** Conceptualization, Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data at the attach file.

Acknowledgements

This publication is part of the R+D+d projects PID2020-112540RB-C41 (AETHER-UMA), PID2020-112540RB-C42 (AETHER-UCLM) and PID2020-112540RB-C44 (AETHER-US): A smart data holistic approach for context-aware data analytics, all of which are funded by MCIN/AEI/10.13039/501100011033/. Also, it has been partially funded by the R&D projects METAMORFOSIS, Spain (US-1381375) from Junta de Andalucía, and ADAGIO, Alarcos' DAta Governance framework and systems generation, Spain (SBPLY/21/180501/000061), funded by the Consejería de Educación, Cultura *y* Deportes of the Junta de Comunidades de Castilla-La Mancha (Spain) and funding for open access charge: Universidad de Málaga / CBUA.

References

[1] N. Gupta, S. Mujumdar, H. Patel, S. Masuda, N. Panwar, S. Bandyopadhyay, S. Mehta, S. Guttula, S. Afzal, R. Sharma Mittal, et al., Data quality for machine learning tasks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 4040–4041.

[2] C. Barba-González, J. García-Nieto, M. del Mar Roldán-García, I. Navas-Delgado, A.J. Nebro, J.F. Aldana-Montes, BIGOWL: Knowledge centered big data analytics, Expert Syst. Appl. 115 (2019) 543–556.

[3] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, V. Munigala, Overview and importance of data quality for machine learning tasks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3561–3562.

[4] H. Challa, N. Niu, R. Johnson, Faulty requirements made valuable: On the role of data quality in deep learning, in: 2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), IEEE, 2020, pp. 61–69.

[5] R. Gu, Y. Qi, T. Wu, Z. Wang, X. Xu, C. Yuan, Y. Huang, SparkDQ: Efficient generic big data quality management on distributed data-parallel computation, J. Parallel Distrib. Comput. 156 (2021) 132–147.

[6] F. Rosner, A. Sorokoumov, Drunken data quality, 2015.

[7] S. Schelter, P. Schmidt, T. Rukat, M. Kiessling, A. Taptunov, F. Biessmann, D. Lange, Deequ-data quality validation for machine learning pipelines, 2018.

[8] Apache, Apache griffin, 2018, [online].

[9] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Vol. 72, Springer, 2015.

[10] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data Sci. J. 14 (2015) 1–10.

[11] N. Paton, Automating data preparation: Can we? should we? must we? in: 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, 2019, pp. 1–6.

[12] F. Chiang, R.J. Miller, Discovering data quality rules, Proc. VLDB Endow. 1 (1) (2008) 1166–1177.

[13] C. Cichy, S. Rass, An overview of data quality frameworks, IEEE Access 7 (2019) 24634–24648.

[14] S.-T. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A. Yeo, A. Talaei-Khoei, Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature, Int. J. Med. Inform. 82 (1) (2013) 10–24.

[15] S.G. Johnson, S. Speedie, G. Simon, V. Kumar, B.L. Westra, A data quality ontology for the secondary use of EHR data, in: AMIA Annual Symposium Proceedings, Vol. 2015, American Medical Informatics Association, 2015, p. 1937.

[16] S.G. Johnson, S. Speedie, G. Simon, V. Kumar, B.L. Westra, Application of an ontology for characterizing data quality for a secondary use of EHR data, Appl. Clin. Inform. 7 (01) (2016) 69–88.

[17] C. Daraio, M. Lenzerini, C. Leporelli, P. Naggar, A. Bonaccorsi, A. Bartolucci, The advantages of an ontology-based data management approach: openness, interoperability and data quality, Scientometrics 108 (1) (2016) 441–455.

[18] F.-B. Mocnik, A. Mobasheri, L. Griesbaum, M. Eckle, C. Jacobs, C. Klonner, A grounding-based ontology of data quality measures, J. Spatial Inf. Sci. 2018 (16) (2018) 1–25.

[19] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Comput. Surv. (CSUR) 41 (3) (2009) 1–52.

[20] K. Kluza, W.T. Adrian, P. Wiśniewski, A. Ligeza, Understanding decision model and notation: DMN research directions and trends, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2019, pp. 787–795.

[21] T. Biard, A.L. Mauff, M. Bigand, J.-P. Bourey, Separation of decision modeling from business process modeling using new "Decision Model and Notation"(DMN) for automating operational decision-making, in: Working Conference on Virtual Enterprises, Springer, 2015, pp. 489–496.

[22] K. Figl, J. Mendling, G. Tokdemir, J. Vanthienen, What we know and what we do not know about DMN, Enterp. Model. Inf. Syst. Archit. (EMISAJ) 13 (2018) 2–21.

[23] C. Corea, J. Blatt, P. Delfmann, A tool for decision logic verification in DMN decision tables, in: BPM (PhD/Demos), 2019, pp. 169–173.

[24] A. Sundararaman, A framework for linking Data Quality to business objectives in decision support systems, in: 3rd International Conference on Trendz in Information Sciences & Computing (TISC2011), IEEE, 2011, pp. 177–181.

[25] Á. Valencia-Parra, L. Parody, Á.J. Varela-Vaca, I. Caballero, M.T. Gómez-López, DMN4DQ: When data quality meets DMN, Decis. Support Syst. 141 (2021) 113450.

[26] D. Calvanese, M. Dumas, Ü. Laurson, F.M. Maggi, M. Montali, I. Teinemaa, Semantics and analysis of DMN decision tables, in: International Conference on Business Process Management, Springer, 2016, pp. 217–233.

[27] I. Horrocks, P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, et al., SWRL: A semantic web rule language combining OWL and RuleML, W3C Memb. Submiss. 21 (79) (2004) 1–31.

[28] N. Noy, D.M. (Hrsg.), Ontology Development 101: A Guide To Creating Your First Ontology, Technical report, Stanford knowledge systems laboratory technical report KSL-01-05 and . . . , 2001.

[29] W. Yun, X. Zhang, Z. Li, H. Liu, M. Han, Knowledge modeling: A survey of processes and techniques, Int. J. Intell. Syst. 36 (4) (2021) 1686–1720.

[30] J. Gelernter, J. Jha, Challenges in ontology evaluation, J. Data Inf. Qual. (JDIQ) 7 (3) (2016) 1–4.

[31] I. Horrocks, P.F. Patel-Schneider, S. Bechhofer, D. Tsarkov, OWL rules: A proposal and prototype implementation, Web Semant.: Sci., Serv. Agents World Wide Web 3 (1) (2005) 23–40.

[32] B.N. Grosof, T.C. Poon, SweetDeal: Representing agent contracts with exceptions using semantic web rules, ontologies, and process descriptions, Int. J. Electr. Commer. 8 (4) (2004) 61–97.

[33] J.A. Khan, S. Kumar, OWL, RDF, RDFS inference derivation using Jena semantic framework & pellet reasoner, in: 2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014), IEEE, 2014, pp. 1–8.

[34] C. Batini, M. Scannapieco, et al., Data and Information Quality, Springer, 2016.

[35] L. Ehrlinger, B. Werth, W. Wöß, Automated continuous data quality measurement with quaiie, Int. J. Adv. Softw. 11 (3) (2018) 400–417.

[36] S. Shrivastava, D. Patel, A. Bhamidipaty, W.M. Gifford, S.A. Siegel, V.S. Ganapavarapu, J.R. Kalagnanam, Dqa: Scalable, automated and interactive data quality advisor, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 2913–2922.

[37] H. Zou, K. Xiang, A novel rigorous measurement model for big data quality characteristics, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 2699–2708.

[38] F. Serra, V. Peralta, A. Marotta, P. Marcel, Use of context in data quality management: a systematic literature review, 2022, pp. 1–40, arXiv preprint arXiv:2204.10655.

[39] R.Y. Wang, A product perspective on total data quality management, Commun. ACM 41 (2) (1998) 58–65.

[40] I.O. for Standardization, Systems and Software Engineering: Systems and Software Quality Requirements and Evaluation (SQuaRE): Measurement of System and Software Product Quality, ISO, 2016.

[41] F. Gualo, M. Rodríguez, J. Verdugo, I. Caballero, M. Piattini, Data quality certification using ISO/IEC 25012: Industrial experiences, J. Syst. Softw. 176 (2021) 110938.

[42] I. Caballero, F. Gualo, M. Rodríguez, M. Piattini, BR4DQ: A methodology for grouping business rules for data quality evaluation, Inf. Syst. 109 (2022) 102058.

[43] I.F. Ilyas, X. Chu, Data Cleaning, Morgan & Claypool, 2019.

[44] I. Caballero, E. Verbo, C. Calero, M. Piattini, A data quality measurement information model based on ISO/IEC 15939, in: ICIQ, Cambridge, MA, 2007, pp. 393–408.

[45] ISO, ISO/IEC/IEEE International Standard - Systems and Software Engineering–Measurement Process, ISO/IEC/IEEE 15939:2017(E), 2017, pp. 1–49, http://dx.doi.org/10.1109/IEEESTD.2017.7907158.

[46] E. Reynares, M.L. Caliusco, M.R. Galli, A set of ontology design patterns for reengineering SBVR statements into OWL/SWRL ontologies, Expert Syst. Appl. 42 (5) (2015) 2680–2690, http://dx.doi.org/10.1016/j.eswa.2014.11.012.

[47] K. Figl, J. Mendling, G. Tokdemir, J. Vanthienen, What we know and what we do not know about DMN, EMISA Forum 38 (1) (2018) 24–26.

[48] A. Paschke, S. Könnecke, RuleML - DMN translator, in: T. Athan, A. Giurca, R. Grütter, M. Proctor, K. Teymourian, W.V. Woensel (Eds.), Supplementary Proceedings of the RuleML 2016 Challenge, Doctoral Consortium and Industry Track Hosted By the 10th International Web Rule Symposium, RuleML 2016, New York, USA, July 6-9, 2016, in: CEUR Workshop Proceedings, 1620, CEUR-WS.org, 2016, pp. 1–13, URL http://ceur-ws.org/Vol-1620/paper4.pdf.

[49] C. Fürber, M. Hepp, Using semantic web technologies for data quality management, in: Handbook of Data Quality: Research and Practice, Springer, 2013, pp. 141–161.

[50] C. Fürber, M. Hepp, Using SPARQL and SPIN for data quality management on the semantic web, in: Business Information Systems: 13th International Conference, BIS 2010, Berlin, Germany, May 3-5, 2010. Proceedings 13, Springer, 2010, pp. 35–46.

[51] C. Fürber, M. Hepp, Swiqa–a semantic web information quality assessment framework, in: ECIS 2011 Proceedings, 2011, pp. 1–8.

[52] C. Fürber, M. Hepp, Towards a vocabulary for data quality management in semantic web architectures, in: Proceedings of the 1st International Workshop on Linked Web Data Management, 2011, pp. 1–8.

[53] V.C. Pezoulas, K.D. Kourou, F. Kalatzis, T.P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, F. Skopouli, S. De Vita, A.G. Tzioufas, et al., Medical data quality assessment: On the development of an automated framework for medical data curation, Comput. Biol. Med. 107 (2019) 270–283.

[54] Y. Liu, Y. Wang, K. Zhou, Y. Yang, Y. Liu, Semantic-aware data quality assessment for image big data, Future Gener. Comput. Syst. 102 (2020) 53–65.

[55] D. Souza, R. Belian, A.C. Salgado, P.A. Tedesco, Towards a context ontology to enhance data integration processes, in: ODBIS, 2008, pp. 49–56.

[56] L. Bertossi, M. Milani, Ontological multidimensional data models and contextual data quality, J. Data Inf. Qual. (JDIQ) 9 (3) (2018) 1–36.

[57] Y.D. Kawtar, L. Hind, C. Dalila, Ontology-based knowledge representation for open government data, Int. J. Intell. Syst. Appl. Eng. 10 (4) (2022) 761–766.

[58] A. Nayak, B. Bozic, L. Longo, (Linked) Data Quality Assessment: An Ontological Approach, Technological University Dublin, 2021.

[59] T. Wang, Y. Zeng, M. Jin, R. Jia, A unified framework for task-driven data quality management, 2021, arXiv preprint arXiv:2106.05484.

[60] R. Almeida, P. Maio, P. Oliveira, J. Barroso, Ontology based rewriting data cleaning operations, in: Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, 2016, pp. 85–88.

[61] J. Debattista, C. Lange, S. Auer, daQ, an ontology for dataset quality information, in: LDOW, 2014, pp. 1–9.

[62] R. Albertoni, A. Isaac, Introducing the data quality vocabulary (DQV), Semant. Web 12 (1) (2021) 81–97.

[63] N. Arruda, J. Alcântara, V. Vidal, A. Brayner, M. Casanova, V. Pequeno, W. Franco, A fuzzy approach for data quality assessment of linked datasets, in: International Conference on Enterprise Information Systems, Vol. 1, SciTePress, 2019, pp. 399–406.

[64] S. Geisler, S. Weber, C. Quix, Ontology-based data quality framework for data stream applications, in: ICIQ, 2011, pp. 1–15.

[65] S. Geisler, C. Quix, S. Weber, M. Jarke, Ontology-based data quality management for data streams, J. Data Inf. Qual. (JDIQ) 7 (4) (2016) 1–34.

[66] F. Bauer, M. Kaltenböck, Linked Open Data: The Essentials, Vol. 710, Edition mono/monochrom, Vienna, 2011, p. 21.

[67] P. Ristoski, H. Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, J. Web Semant. 36 (2016) 1–22.

[68] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semant. Web 7 (1) (2016) 63–93.

[69] D. Wienand, H. Paulheim, Detecting incorrect numerical data in dbpedia, in: The Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings 11, Springer, 2014, pp. 504–518.

[70] P.N. Mendes, H. Mühleisen, C. Bizer, Sieve: linked data quality assessment and fusion, in: Proceedings of the 2012 Joint EDBT/ICDT Workshops, 2012, pp. 116–123.

[71] E. Rahm, H.H. Do, et al., Data cleaning: Problems and current approaches, IEEE Data Eng. Bull. 23 (4) (2000) 3–13.

[72] D. Calvanese, M. Dumas, Ü. Laurson, F.M. Maggi, M. Montali, I. Teinemaa, Semantics and analysis of DMN decision tables, in: M. La Rosa, P. Loos, O. Pastor (Eds.), Business Process Management, Springer International Publishing, 2016, pp. 217–233.

[73] J. Poore, T. Nemecek, Reducing food's environmental impacts through producers and consumers, Science 360 (6392) (2018) 987–992.

[74] Á. Valencia Parra, Analysis of big data architectures and pipelines: Challenges and opportunities, in: Máster Universitario en Ingeniería Informática, 2019, pp. 1–96.

[75] J. de Haro-Olmo, Á. Valencia-Parra, Á.J. Varela-Vaca, J.A. Álvarez-Bermejo, M.T. Gómez-López, ELI: an IoT-aware big data pipeline with data curation and data quality, PeerJ Comput. Sci. 141 (2023) 113450.

[76] C. Schader, M. Curran, A. Heidenreich, J. Landert, J. Blockeel, L. Baumgart, B. Ssebunya, S. Moakes, S. Marton, G. Lazzarini, U. Niggli, M. Stolze, Accounting for uncertainty in multi-criteria sustainability assessments at the farm level: Improving the robustness of the SMART-Farm Tool, Ecol. Indic. 106 (2019) 105503.

[77] E. Jerhamre, C.J.C. Carlberg, V. van Zoest, Exploring the susceptibility of smart farming: Identified opportunities and challenges, Smart Agric. Technol. 2 (2022) 100026, http://dx.doi.org/10.1016/j.atech.2021.100026, URL https://www.sciencedirect.com/science/article/pii/S2772375521000265.

[78] C. Schader, L. Baumgart, J. Landert, A. Muller, B. Ssebunya, J. Blockeel, R. Weisshaidinger, R. Petrasek, D. Mészáros, S. Padel, C. Gerrard, L. Smith, T. Lindenthal, U. Niggli, M. Stolze, Using the sustainability monitoring and assessment routine (SMART) for the systematic analysis of trade-offs and synergies between sustainability dimensions and themes at farm level, Sustainability 8 (3) (2016) 1–20.

[79] W.G. de Almeida, R.T. de Sousa, F.E. de Deus, G. Daniel Amvame Nze, F.L.L. de Mendonça, Taxonomy of data quality problems in multidimensional Data Warehouse models, in: 2013 8th Iberian Conference on Information Systems and Technologies (CISTI), 2013, pp. 1–7.

[80] Á. Valencia Parra, Á.J. Varela Vaca, M.T. Gómez López, P. Ceravolo, CHAMA-LEON: framework to improve data wrangling with complex data, in: ICIS 2019: 4th International Conference on Information Systems (2019), Association for Information Systems (AIS), 2019, pp. 1–17.

[81] A. Even, G. Shankaranarayanan, Utility-driven assessment of data quality, Data Base 38 (2) (2007) 75–93, http://dx.doi.org/10.1145/1240616.1240623.

[82] M. Janssen, H. van der Voort, A. Wahyudi, Factors influencing big data decision-making quality, J. Bus. Res. 70 (2017) 338–345.

[83] United States Department of Agriculture, Data from: A field-scale sensor network data set for monitoring and modeling the spatial and temporal variation of soil moisture in a dryland agricultural field, 2007, https://agris.fao.org/agris-search/search.do?recordID=US2019X00214, Last accessed on 2021-06-02.

[84] C.K. Gasch, D.J. Brown, C.S. Campbell, D.R. Cobos, E.S. Brooks, M. Chahal, M. Poggio, A field-scale sensor network data set for monitoring and modeling the spatial and temporal variation of soil water content in a dryland agricultural field, Water Resour. Res. 53 (12) (2017) 10878–10887, http://dx.doi.org/10.1002/2017WR021307, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017WR021307, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR021307.

[85] F.J. de Haro-Olmo, Á. Valencia-Parra, Á.J. Varela-Vaca, J.A. Álvarez-Bermejo, Data curation in the Internet of Things: A decision model approach, Comput. Math. Methods (2021) e1191.

[86] C.K. Gasch, D.J. Brown, E.S. Brooks, M. Yourek, M. Poggio, D.R. Cobos, C.S. Campbell, A pragmatic, automated approach for retroactive calibration of soil moisture sensors using a two-step, soil-specific correction, Comput. Electron. Agric. 137 (2017) 29–40, http://dx.doi.org/10.1016/j.compag.2017.03.018, URL https://www.sciencedirect.com/science/article/pii/S0168169916304288.

[87] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, Experimentation in Software Engineering, Springer, 2012, http://dx.doi.org/10.1007/978-3-642-29044-2.

[88] R.J. Wieringa, Design Science Methodology for Information Systems and Software Engineering, Springer, 2014.

[89] A. Benítez-Hidalgo, C. Barba-González, J. García-Nieto, P. Gutiérrez-Moncayo, M. Paneque, A.J. Nebro, M. del Mar Roldán García, J.F.A. Montes, I.N. Delgado, TITAN: A knowledge-based platform for Big Data workflow management, Knowl.-Based Syst. 232 (2021) 107489, http://dx.doi.org/10.1016/j.knosys.2021.107489.