Article

A FAIR Perspective on Data Quality Frameworks

Nicholas Nicholson [1,*,†], **Raquel Negrao Carvalho** [1] **and Iztok Štotl** [2,3,†]

1 European Commission, Joint Research Centre (JRC), I-21027 Ispra, Italy; raquel.carvalho@ec.europa.eu
2 Department of Endocrinology, Diabetes and Metabolic Diseases, University Medical Centre Ljubljana, 1000 Ljubljana, Slovenia; iztok.stotl@guest.arnes.si
3 Faculty of Medicine, University of Ljubljana, 1000 Ljubljana, Slovenia
* Correspondence: nicholas.nicholson@ec.europa.eu; Tel.: +39-0332-78-9365
† These authors contributed equally to this work.

Abstract

Despite considerable effort and analysis over the last two to three decades, no integrated scenario yet exists for data quality frameworks. Currently, the choice is between several frameworks dependent upon the type and use of data. While the frameworks are appropriate to their specific purposes, they are generally prescriptive of the quality dimensions they prescribe. We reappraise the basis for measuring data quality by laying out a concept for a framework that addresses data quality from the foundational basis of the FAIR data guiding principles. We advocate for a federated data contextualisation framework able to handle the FAIR-related quality dimensions in the general data contextualisation descriptions and the remaining intrinsic data quality dimensions in associated dedicated context spaces without being overly prescriptive. A framework designed along these lines provides several advantages, not least of which is its ability to encapsulate most other data quality frameworks. Moreover, by contextualising data according to the FAIR data principles, many subjective quality measures are managed automatically and can even be quantified to a degree, whereas objective intrinsic quality measures can be handled to any level of granularity for any data type. This serves to avoid blurring quality dimensions between the data and the data application perspectives as well as to support data quality provenance by providing traceability over a chain of data processing operations. We show by example how some of these concepts can be implemented at a practical level.

Keywords: data quality frameworks; FAIR data principles; data contextualisation; metadata; quality provenance; data pathway; knowledge management; federated data

1. Introduction

The volume of peer-reviewed literature and other technical documentation over the previous two to three decades proposing various quality metrics for data testifies to the elusiveness of convergence towards an all-embracing data quality (DQ) solution. Not only do the different domains of business, environment, health, big data, sensor networks (Internet of Things), etc., approach data quality with a different set of perspectives but data can be described on a wide spectrum from atomic data elements to very large, distributed databases as well as to dynamic data flows and, more recently, to large language models (LLMs). Moreover, the needs of different data applications using the same sets of data may widely differ in terms of the importance they give to specific quality dimensions.

Whereas convergence has generally been achieved for structured data on the more intrinsic DQ dimensions [1,2], agreement on an overall DQ framework has been more

problematic, despite the serviceability of individual frameworks to their intended purposes. Indeed, as borne out by several extensive reviews (discussed in Section 2), selecting an appropriate DQ framework for a particular application is not straightforward since each framework imposes its own set of constraints and is generally prescriptive of the quality dimensions it requires. Real-time data flows, machine learning, and federated data fusion applications impose yet further sets of unique constraints [3–7]. Our own particular use case concerns the need to determine quality indices of diabetes-related health indicators. Notwithstanding the specific reference to diabetes and health, the use case is pertinent to all types of indicators in other domains and serves to illustrate the limitations of currently available DQ frameworks.

The way in which an indicator is derived generally follows an inter-linked chain of data capture and data processing steps that together play a critical role in the interpretation and usability of the final indicator. Without knowing the assumptions, limitations, and bias introduced at each stage throughout the data chain, the indicator cannot be used with any degree of certainty. This prompts the need to record such information for later retrieval.

Moreover, this chaining process will involve most of the individual data entity types identified by Haug [8], which are elaborated under point 2 of the enumerated list in Section 2. For the indicator quality use case, a DQ paradigm is needed for a composite process that works across all these data entity types without having to implement several different frameworks.

Ideally, the paradigm would offer within a single framework the means to:

1. Contextualise any data entity with standardised metadata elements allowing data users to drill into the context to the degree of granularity necessary to understand the applicability of the data for their specific needs.
2. Describe the intrinsic DQ dimensions used for a given data type, the definition and assessment of which could draw from standard definitions without being prescriptive to a rigid framework defined by a fixed number of dimensions and precise definitions.
3. Describe the quantitative/qualitative metrics used for measuring/estimating those dimensions, which could draw from standard definitions, yet also allow individual tailoring to the needs.
4. Provide a level of mapping between tailor-made quality definitions/metrics and standard ones.
5. Link to standard definitions (to avoid reproducing them within the framework) to maintain integrity with the definitions.
6. Link data derived in a processing chain with data from earlier steps in the chain to derive composite quality metrics with a full degree of flexibility, enabling data users to drill into the composite measures to obtain the more granular quality metric information.
7. Query the contextual and intrinsic quality information and/or links to parts of the processing chain on a need-to-know basis allowing users or AI applications to ascertain the fitness-for-use of the associated data for a given application.

Notwithstanding the efforts described in Section 1.1 to categorise DQ dimensions into a comprehensive taxonomy that could accommodate most DQ frameworks, many of the enumerated needs discussed above remain largely unaddressed.

1.1. Generic DQ Model

A generic DQ model has been proposed to accommodate the different aspects and points of focus of any DQ framework [9]. The model builds on the ISO 25012 data quality model for software product quality requirements and evaluation (SquaRE) standard [2] and underlines the need to classify different quality dimensions under separate conceptual categories. The authors illustrate the quality dimensions on a quality wheel categorised

within a hierarchy of three levels, the top two levels of which are described in Table 1. The third level contains many other terms, incorporating all the terms identified in the other reviews.

Table 1. The top-level and second-level categories of quality dimensions described by the data quality model after Miller et al. [9].

Top-Level Category	Second-Level Category
Contextual	Accessibility Confidentiality Efficiency Governance Compliance Traceability Precision Understandability Usefulness
System-dependent	Recoverability Portability Semantics Availability Quantity
Inherent	Credibility Completeness Consistency Accuracy Currentness

The authors consider the categories as context-specific since they include both inherent and system-dependent characteristics. This is an important insight that avoids the attenuation of DQ indices caused by combining quality dimensions not directly relevant to each other. For instance, since the DQ dimension "accessibility" has little relevance to "completeness", it is classified under a different top-level category.

1.2. Data-Centric View of Data Quality Versus DQ-Centric View of Data

Whereas the fundamental reason for wanting to measure data quality is arguably the need to use or reuse of data for some specific purpose, the data are independent of that need and should ideally be measured separately from it. However, many DQ frameworks introduce quality dimensions such as understandability, usefulness, interpretability, and relevance that are subjective measures of the data user from an application point of view. This consideration underlies two different approaches to viewing data quality.

One can either take a primarily DQ-centric view of data, which is the concept behind most DQ frameworks, or a predominantly data-centric view of data quality, in which quality is ascertained from a holistic description of the actual data. The latter approach aligns closely with the philosophy behind the FAIR data model [10], which provides several advantages.

1.3. FAIR Data Principles

FAIR is an acronym describing four foundational principles of data relating to findability, accessibility, interoperability, and reusability. DQ-centric frameworks treat these terms as individual quality dimensions at various category hierarchical levels. In the model of [9] for instance, "findability" is classified as a subdivision of the category "accessibility", and "interoperability" and "reusability" are classified together under the category "useful-

ness". In contrast, the FAIR data model considers data quality as a function of all the FAIR data principles. Moreover, although the principles do not explicitly prescribe how quality should be measured, they view quality primarily as a data application concern that can be addressed from the complementarity of the data's metadata for determining the suitability of the data for the required needs [11]. Consequently, a set of quality labels need only be attributed once to the data rather than many times dependent upon the differing data application DQ requirements. The latter need only be matched with the DQ labels of the data to understand the suitability of the data for the required purpose. This has the further advantage of allowing the data application to specify its own specific minimal information standards (such as minimum information for biological and biomedical investigations, MIBBI, or minimum information about a microarray experiment, MIAME, for biological and biomedical investigations) without their necessarily being an explicit part of the DQ model itself [11]. Conformance to such specific information standards can be assessed from the general contextual content of the data source.

The purpose of the work we present here is fivefold, namely (i) to lay out a concept for a DQ framework that addresses data quality from the foundational basis of the FAIR data guiding principles such that it places these principles central to the framework rather than subjugating them to various hierarchical levels of DQ dimensions; (ii) to provide within the framework a flexible and scalable way of incorporating DQ dimensions, taking into account the different needs of different data entity types and allowing them to associate with different metrics; (iii) to show by example how some of these concepts can be implemented at a practical level by adding some extensions to an existing general data contextualisation framework; (iv) to illustrate how such a framework can encapsulate fundamental aspects of many other frameworks where relevant to the data/process needs; and (v) to stimulate further discussion and research into data contextualisation models that can improve and extend the ideas presented here.

The primary focus of this study is on a generic description framework. It does not consider the DQ assessment and improvement strategies that other DQ frameworks employ since these can be applied regardless within the agreements of a given data domain and referenced from the framework as described in Section 3. It also does not define any data quality metrics since these will be needs-specific and the framework can reference the metrics models required for a given purpose.

2. Materials and Methods

In our quest to find an existing DQ model able to fulfil the needs of the use case introduced in the Introduction, we looked at several comprehensive, recent reviews of DQ frameworks [8,12–15].

Two of these reviews [8,12] acknowledged the difficulty of finding a single, over-arching framework due to all the variables in play ranging from data and data-structure type, application needs, subjectiveness of DQ dimension, and domain area. The reviews categorised the commonly used frameworks to help researchers identify the most suitable one for their particular data needs. Cichy and Rass [12] illustrated on a quality wheel the 20 quality dimensions they found common to more than one of the DQ frameworks reviewed. Haug [8] noted that data quality is a multidimensional concept that can be considered a set of DQ dimensions each describing a particular characteristic of data quality, sometimes grouped under DQ categories. Haug presented his analysis in terms of DQ classifications. The frameworks reviewed by Cichy and Rass are listed in Table 2. Three of these (TDQM, DQA, and AIMQ) featured in the review of Haug who also analysed a number of other schema, including novel combinations of DQ dimensions.

Table 2. Commonly used DQ frameworks analysed by Cichy and Rass [12]. Italicised text refers to the same framework analysed also by Haug [8].

Commonly Used DQ Frameworks Reviewed
AIM quality (AIMQ) [16]
Comprehensive methodology for data quality management (CDQ) [17]
Cost-effect of low data quality (COLDQ) [18]
Data quality assessment (DQA) [19]
Data quality assessment framework (DQAF) [20]
A data quality practical approach [21]
Data quality methodology for heterogeneous data (HDQM) [22]
Hybrid information quality management (HIQM) [23]
Observe–orient–decide–act methodology for data quality (OODA DQ) [24]
Task-based data quality method (TBDQ) [25]
Total data quality management (TDQM) [26]
Total information quality management (TIQM) [27]

Although the reviews had different points of focus and different aims, their findings are closely aligned in relation to the major issues encountered in the DQ frameworks analysed, namely:

1. Data quality is a multidimensional concept but there is little agreement on the DQ dimensions it should comprise.

2. DQ dimensions are, in general, highly context dependent and vary with the type of categorisation under consideration (such as access to data or data interpretability). In addition, DQ dimensions are dependent upon the nature of the data entity (e.g., data item, data field, data record, dataset, database, database collection, etc.) for which a DQ dimension such as completeness may be specified in different terms.

3. DQ frameworks differ widely in the number of DQ dimensions they define (ranging from a few to more than thirty). For structured data, commonly agreed DQ dimensions are completeness, accuracy, and timeliness, followed by consistency and accessibility. The relevance of all these dimensions and the exact meaning are however dependent on the nature of the data entity. Moreover, structured data and unstructured data may require quite different DQ dimensions (interpretability and conciseness, for example, are more relevant for unstructured data).

4. DQ dimensions can have objective and subjective measures or a mix of both. Both measures have their uses; some dimensions are either difficult to score quantitatively or can only be performed so in specific data-application terms (in particular for relevance- or presentational-type dimensions). The rigour of a measure (fitness-for-use) is also data-application specific.

5. Whereas DQ frameworks represent different areas of application, widely different frameworks can be used within the same application area.

6. DQ dimensions often overlap and do not have consistent interpretations resulting in lack of clarity of how to map or measure quality dimensions in practical implementations. Moreover, the processes for assessing data quality differ significantly depending on the mix of subjective/objective measures, the level of data granularity, and the nature of the data (whether primary or derived).

7. Ensuring orthogonality of selected dimensions is important to avoid these overlapping meanings and to ensure distinguishability.

Haug [8] made several recommendations for future DQ development initiatives that included the need to clarify the focus of application, establish the relevant DQ dimensions within distinguishable categories, specify the evaluation perspectives, make explicit

the data entities and data structure, and clearly define the dimensions—especially with consideration to existing definitions.

Other reviews in the realm of health data, examined DQ assessment methods and provided useful cross matrices of assessment methods with various DQ dimensions [13,14]. Bian et al. [13] analysed 3 reviews, 20 frameworks, and 226 DQ studies; the authors extracted 14 DQ dimensions and 10 assessment methods. Declerck et al. [14] analysed 22 reviews comprising 22 frameworks, 23 DQ dimensions, 62 definitions of DQ dimensions, and 38 assessment methods; the authors mapped all the DQ dimensions to seven of the nine dimensions of the data quality framework of the European Institute for Innovation through Health Data. These reviews further highlight the lack of consensus regarding the terminology, definition, and assessment methods for DQ dimensions. In consequence, Declerck et al. [14] recommended that future research should shift its focus toward defining and developing specific DQ requirements tailored to each use case rather than pursuing an elusive quest for a rigid framework defined by a fixed number of dimensions and precise definitions. The authors highlighted the need to complete the collection of aspects within each quality dimension and elaborated a full set of assessment methods.

One further review compared the different data quality frameworks, listed in Table 3, underpinned by regulatory, statutory, or governmental standards and regulations in various domains [15].

Table 3. Data quality frameworks underpinned by regulatory, statutory, or governmental standards and regulations in various domains, analysed by Miller et al. [10].

Regulatory/Statutory/Governmental DQ Frameworks Reviewed
TDQM [26]
ISO 8000 [28]
ISO 25012 [2]
Fair information practice principles (FIPPS) [29]
Quality assurance framework of the European statistical system (ESS QAF) [30]
UK government data quality framework [31]
Data management body of knowledge (DAMA DMBoK) [32]
IMF data quality assessment framework (DQAF) [33]
Basel Committee on banking supervision standard (BCBS 239) [34]
ALCOA+ principles [35]
WHO data quality assurance (DQA) framework [36]

This review also highlighted the prominence given to the quality dimensions of accuracy, completeness, consistency, and timeliness but noted the evolution of their meanings across different domains. The authors emphasised the need for newer and more modern quality dimensions to be recognised and integrated into all-purpose DQ frameworks to keep pace with the needs of emerging technologies, such as AI systems based on LLMs.

A further issue relates to the "pervasive DQ problem", highlighted by Karr et al. [37], regarding the failure to distinguish clearly between original data attributes and derived attributes calculated from original data, as well as to differentiate between the three hyperdimensions of data, process, and use. Regarding the latter, Chen et al. [38] observed inadequate attention had been given to process and use. For the indicator-derivation use case, understanding the quality of the processes involved has a major influence on the quality of the resulting indicator.

2.1. Rationalisation of Quality Dimensions in a General Data Contextualisation Schema

Given these limitations of current DQ frameworks and to appreciate the ways in which a data-holistic framework can address them, we start by considering the quality dimension

wheel of Miller et al. [9], due both to the comprehensiveness of the dimensions captured from many DQ frameworks and the hierarchical categorisation consequently proposed by the authors (c.f. Table 1 for the first two levels of categorisation). The third level of categorisation contains more than 240 terms, some of which are relatively synonymous (e.g., "usability" and "useableness" in the Usefulness category) and others of which are characterised by different levels of abstraction (e.g., "metadata" and "original" in the Traceability category).

We next make a distinction between the quality dimensions relating more specifically to data-application needs that are difficult to quantify generally and those that are objective to the data themselves that are more amenable to quantification. The results are summarised in Table 4, where the first column refers to data-application quality dimensions and the second column to data-intrinsic quality dimensions. We have added into the second column the relevant third-level terms that were before categorised under the second-level data-application oriented quality dimensions. The first three terms in column 1 of the table will generally have different metrics for different applications depending on the specific needs of those applications. The fourth term "Quantity" could be argued as being intrinsic to the data; however, the amount of data will be known from the general data contextualisation attributes, and it is then an application requirement to understand if the quantity of the data described is appropriate to its needs. We therefore treat it as a data-application dimension.

Table 4. Separation of the data-application types of DQ dimensions (left-hand column) from the DQ model of Miller et al [9] and the re-inclusion of third-level terms that can be considered more objective of the data themselves (right-hand column).

Second-Level Data Application DQ Dimensions	Associated Third-Level DQ Dimensions That Can Be Considered Objective of the Data Themselves
Usefulness	Artificiality Expandability Interoperability Naturalness Reusability Uniqueness
Understandability	Clean Format Information-to-noise ratio Presentation Semiotic Structure Transformation Visualisation
Efficiency	Minimality
Quantity	Coverage Data volume

For the remaining dimensions, we then identify the quality dimensions that can be said to form part of a data entity's general contextualisation information, as elaborated under the FAIR data principle sub-levels [10]: F.2—"data are described with rich metadata"; A.2—"metadata are accessible"; I.1—"(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation"; I.2—"(meta)data use vocabularies that follow FAIR principles"; I.3—"(meta)data include qualified references to other (meta)data"; and R.1—"(meta)data are richly described with a plurality of relevant and

accurate attributes". Table 5 lists the associated second-level quality dimensions together with the FAIR model sub-levels to which they relate.

Table 5. The quality dimensions that can be considered as forming an inherent part of the formal contextual description of the data, as elaborated under the FAIR data principles.

Quality Dimensions Included from the Second-Level or Third-Level Components of the DQ Model
Traceability (FAIR data principle I.3)
Availability (FAIR data principle A.2)
Semantics (FAIR data principle I.1, I.2, R.1)
Accessibility (FAIR data principle A.2)
Currentness (FAIR data principle F.2)

It should be noted that also in this case some of the third-level terms categorised under the second-level terms in Table 5 contain both application-specific dimensions and data-specific dimensions. For example, Traceability includes application-specific dimensions, such as "quality of methodology" and "translatability" as well as data-specific dimensions, such as "source", "provenance", and "documentation".

After the completion of these two steps, the second-level, data-intrinsic categories that do not directly form part of the general data contextualisation attributes are shown in Table 6.

Table 6. The remaining second-level data-intrinsic categories that do not directly form part of the general data contextualisation attributes.

Quality Dimensions Included from the Second-Level or Third-Level Components of the DQ Model
Governance
Portability
Recoverability
Confidentiality
Compliance
Credibility
Completeness
Consistency
Accuracy
Precision

2.2. Refactoring the Quality Dimensions in a Data Contextualisation Framework

On the premise that data-application-specific dimensions can be ascertained from appropriately contextualised data, the dimensions listed in Tables 5 and 6 may therefore be considered as forming the essential main complement of DQ dimensions in a data contextualisation framework (with perhaps the addition of the third-level term "security", due to its importance for most data domains).

However, given that the dimensions in Table 5 can be factored into the general contextual information described under FAIR, we re-categorise the remaining dimensions of Table 6 under four specific descriptive context spaces shown in Table 7. This allows them to be developed in a relatively modular fashion independently of the other context spaces.

Table 7. Association of the main DQ dimensions to context spaces within a data contextualisation framework.

Data-Specific Context Spaces	Main Level DQ Dimensions to Be Contextualised in a Data Contextualisation Framework
Data Governance Context	Governance Compliance Portability Recoverability
Data Security Context	Security
Data Privacy Context	Confidentiality Compliance
Data Quality Context	Credibility Completeness Consistency Accuracy

Figure 1 illustrates the context spaces and shows the separation of concerns between describing a data entity and establishing the minimum requirements for a data entity by a given data application.
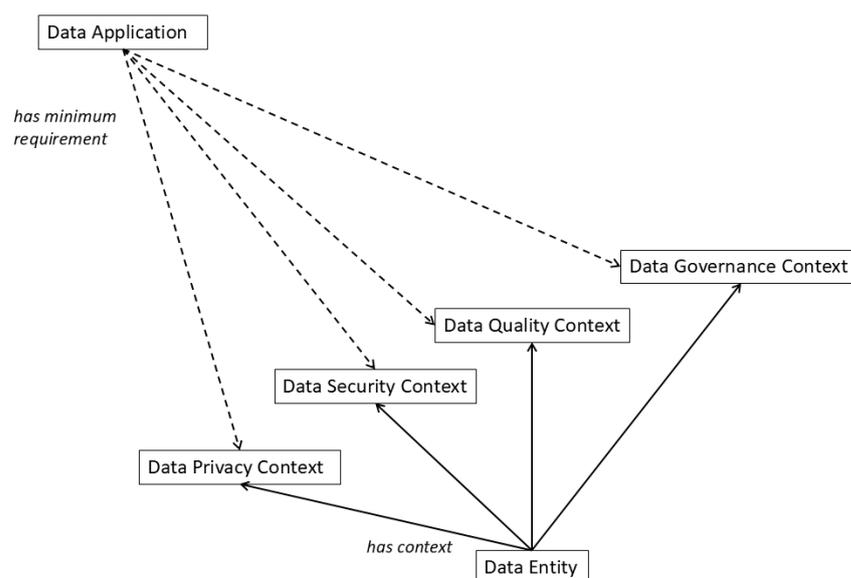

Figure 1. Network graph showing the four specific DQ context spaces directly associated with a data entity and the separation of the data entity's DQ context from that of the data application.

Having rationalised the overall number of necessary DQ dimensions and elevated the FAIR data principles to be the major driver in the general data context space, the next task is to show how this concept can lead to the flexibility allowing an integration of DQ descriptions within a single framework.

3. Results

In this section, we concentrate our attention on how the context spaces of Table 7 can be implemented in practice and consider a few examples only of the general contextualisation attributes to provide an indication of how they can serve the DQ needs of data applications since these are likely to be highly application-specific. We first demonstrate the architectural aspects in Section 3.1 and then implement them in practice in an example in Sections 3.2–3.4. In Section 3.5, we describe how the contextual information can be accessed,

and in Section 3.6, we show how the ideas presented in the example can be used to address application-related DQ needs. Finally, in Section 3.7, we summarise how the concepts can be used to tackle the exigencies of the indictor use case described in the Introduction. It is not our intention to be comprehensive, but merely to give a taste of the flexibility and versatility of the approach and to show how this flexibility can address the requirements of the indicator use case.

3.1. Implementing the Quality Dimensions in a Data Contextualisation Framework

Despite the explosion of references to the FAIR data guiding principles, we could not find a generic data contextualisation framework apart from the semantic ontology-labelled indicator contextualisation integrative taxonomy (SOLICIT) [39], which although mainly addressed to contextualising indicators can in fact be used to contextualise any type of data.

SOLICIT is a generic ontology framework that can be extended at the domain level. It uses standard metadata constructs according to the metamodel of ISO/IEC 11179 metadata registry standard [40]. It also inherits from the mid-level common core ontologies (CCO) [41] and therefore benefits from many standard general classes and relationships already defined. CCO are written in the web ontology language (OWL). Whereas the structure behind a data contextualisation framework is not critical if it supports the essential relations and provides the necessary flexibility, the foundational element of an ontology is extremely useful for defining relationships and concepts. Moreover, the ability to draw on a well-established metadata standard aids interoperability and promotes scalability via reuse of metadata terms.

In SOLICIT, the general data context space is provided by a set of semantic relations and ontology classes which are used to provide contextual elements such as definitions, descriptions, units of measurement/coding schemes, time periods, geospatial locations, derivation methods, and links to any up-stream data sources used in the derivation of a data entity. The operations are discussed at greater length in [39]. These relations can be used to describe the type of information required to determine the quality dimensions of Table 5 as outlined in Section 3.6.

SOLICIT provides the semantic relation "has context" that can be used to associate the context spaces of Table 7. To illustrate this, we consider the "data quality context space". For the sake of argument, we retain four of the five inherent quality dimensions of the model described in Section 1.1 (completeness, accuracy, consistency, and credibility—the fifth quality dimension "currentness" is more appropriately handled in the general contextualisation elements). We do not necessarily need to restrict ourselves only to these four dimensions or even be required to use them; SOLICIT allows us to add others at the extended domain level and use only those dimensions that are necessary.

3.2. Data Quality Context

Figure 2 demonstrates how SOLICIT can be developed to incorporate the data quality context space and quality dimensions. The components in the figure are coloured and prefixed by an identifier referring to the ontology source in which it is curated (BFO, CCO; IEO, AO refer to foundational ontology components and are depicted in orange font; ISO 11179 refers to SOLICIT's ISO/IEC-11179-based components, depicted in black font, and SLCT refers to SOLICIT's own components, depicted in purple font). The prefix DQ refers to classes (in blue font) added to SOLICIT to incorporate the quality aspects discussed in Section 2.

Figure 2. Network graph showing the attributes and relations of the data quality context class. The graph has been elaborated for the "completeness" quality dimension. All components are coloured and preceded by an identifier to identify the source of curation.

The "data quality context" class is created as a subclass of the "context space" class which is itself a subclass of SOLICIT's "context specifying data element" class. The "data quality context" class is linked via the foundational ontology relation "has continuant part" to the individual main dimensions of the space. The main dimensions form the container for the quality indices of the associated sub-dimensions as illustrated in the figure for "completeness", where we have indicated one of the possible sub-dimensions referring to "population completeness" (i.e., a dimension to measure the extent to which data captures all the relevant entities in a defined population).

In the example given, the value of the "population completeness quality index" is referenced by a foundational ontology annotation property "is tokenized by", which is a shorthand means of representing the values of individuals without tracking provenance [42]. The alternative would be to create an instance of the CCO class "information bearing entity" that would contain the value and could also point to the unit of measure.

The method/algorithm for calculating the value is provided using the relation "is input to" referencing a CCO "directive information content entity" or, by consequence, one of its subclasses as shown in Figure 2 (in this case an "algorithm" class). The calculation method here points to a URL that could reference a standard method for calculating a population completeness quality index. The algorithm might itself be expressed in terms of a standard template.

There is in essence no limit to how deep the subclassing may go, which permits considerable flexibility in realising more specific definitions of the quality dimension in cases where such granularity is required. At any given level, the metrics of each individual quality metric can be rolled up into the quality index of the parent level using an algorithm referenced at the parent level. This is shown in the figure for the "population completeness quality index" and its parent "completeness quality index", where the latter also references a value and a calculation algorithm. The roll-up of quality indices can extend to the top "data quality context" level. Figure 2 includes a further attribute of the "has continuant part" (dotted-dashed line) pointing to the "quality index" class to illustrate this possibility.

In rolling up the quality indices to the main dimension level, it is important to ensure the top-level quality dimensions are orthogonal to each other (i.e., independent) to avoid cross-interference and allow the calculation of an independent quality label on each of the

four dimensions (c.f. point 7 of the enumerated list provided in Section 2). This requirement need not necessarily apply to the sub-dimensions since they can be given proportional weightings with respect to their contribution to each of the top-level orthogonal dimensions into which they feed. To achieve orthogonality in practice may not be easy, given the different and overlapping interpretations of DQ dimensions. However, certain measures can be taken to ensure delineation of terms through clear and explicit definitions of what the dimensions include (and what they do not include in cases of ambiguity). Whereas three of the four inherent DQ dimensions considered have good separation, there could potentially be crosstalk between credibility and the other dimensions. The definition of credibility needs therefore to be steered in meaning towards retaining independence from the other dimensions. Thus, for example, a data entity may be deemed accurate since it can be verified that a value is correct (accuracy dimension); however, the means of verifying the value may not be optimal and thus the credibility dimension will capture this fact.

All the other main level-dimensions can be implemented in a similar way and other main level dimensions can be defined at the domain or sub-domain level. Conversely, not all dimensions need to be instanced where they are not relevant—Figure 2 applies just to one data entity type, which is therefore free to specify and define the relevant quality dimensions and sub-dimensions as well as the metrics for measuring them. Furthermore, since a data entity in SOLICIT can point to the data element(s) from which it is derived [39], the whole set of quality contexts become available across the data-process chain.

Moreover, given that a data entity is associated with a derivation process (c.f. Figure 3 of [39]), the process itself can also be associated with a quality context in a similar way, allowing the process itself to be described with the appropriate quality dimensions and thereby addressing the issue raised by Karr et al. [37], discussed in Section 2. Thus, for example, a process involving the integration of datasets could reference the data-cleaning method of Corrales et al. [43] addressing the issues of noise (encapsulating quality dimensions associated with missing values, outliers, high dimensionality (number of variables), imbalanced class, mislabelled class, and duplicate instances. These individual dimensions would from subdimensions of the "noise" main dimension associated with the process quality context and scored according to the associated metrics.

Figure 3. Example of the contextualisation of the aggregated dataset described by the entity "hba1cSet", in terms of the syntax used by the Protégé OWL editor. The bottom highlighted line results from the reasoning process described in Section 3.6.

3.3. Practical Example

In this section, we demonstrate how some of the concepts described in the previous section can be realised by way of a practical example. The example we show (Figure 3) is of a rudimentary (i.e., not comprehensively described) diabetes indicator describing a dataset of common data elements (CDEs), which provides aggregate results within a given population of the haemoglobin A1C (HbA1C) stratified by sex, age, diabetes type, and percentage value. HbA1c is an important blood test for assessing long-term metabolic control in patients with diabetes [44]. The dataset is derived from a dataset containing the individual readings. For this example, the domain refers to non-communicable diseases (NCDs) and diabetes is a sub-domain. SOLICIT would therefore be extended to include the NCD common components and the NCD extension would be extended further in the diabetes domain to include the diabetes-specific components. Notwithstanding the example's relation to health and the consequence that the quality requirements may not be relevant for other domains, it is merely intended to show how quality may be defined in a data contextualisation framework.

In Figure 3, the left-hand column describes the types (essentially OWL classes) of the instance ("hba1cSet") of a SOLICIT Indicator class that describes the dataset. The right-hand column describes the relations to other entities along the lines of Figure 2.

The contextualisation process is enumerated in the following steps:

1. Create an individual (class instance) that has some sub-type of SOLICIT's "context specifying data element" class, which will be the container of the contextualisation information (in the example—hba1cSet).
2. Create the complex class types for standard metadata components described in the ontology using the ISO/IEC 11179 constructs (e.g., the stratification variables and the ISO/IEC 11179 components associated with the individual).
3. Create a set of individuals to instance the various data quality classes, e.g., the relevant contexts that have to be described (in our example, the data quality context, the data governance context, and data privacy context) and reference them to the individuals describing the associated quality dimensions (c.f. step 4).
4. Create the individuals for each quality dimension and reference them to individuals describing the associated quality indices (c.f. step 5). The quality indices can reference the quality indices of the upstream dataset from which the current dataset is derived.
5. Create the individuals describing the algorithm for calculating the quality metrics for a given dimension.
6. Steps 4 and 5 can be repeated to any level of sub-dimension required.
7. Create the remainder of the general contextualisation components, e.g., descriptions, processes, data-extraction scripts, etc.
8. Create a reference to the upstream dataset where relevant (hba1cRaw in the example).

Figure 4 shows an OntoGraf image of the complete set of relations between the various individuals referenced by the example indicator in Figure 3 (class relations are not shown to preserve clarity in the figure). The individuals at the top of the figure refer to general contextualisation elements (upstream dataset, data extraction process, temporal and spatial information). The middle three elements refer to the context spaces, where the data quality context space has been amplified for the completeness dimension and its associated quality index. The context spaces all reference "quality index specification" instances that in turn reference algorithm instances defining the metrics to apply for a composite quality index based on the individual quality indices of the sub-dimensions. Individual quality indexes are specified at the lowest sub-dimension level; for example the three entities in the penultimate row of the figure depict instances of a "population completeness quality index", a "record completeness quality index", and a "data integration completeness quality index",

which each have specific algorithms for determining the respective dimensions (where record completeness signifies the percentage of complete CDE records, and data-integration completeness signifies completeness after integrating other datasets). The composite value for the total completeness quality index is calculated taking all these indices into account and applying the appropriate algorithm (which is referenced by the parent completeness quality index entity). Since population completeness derives from the upstream dataset, the quality index of the current dataset references the quality index of the upstream dataset.
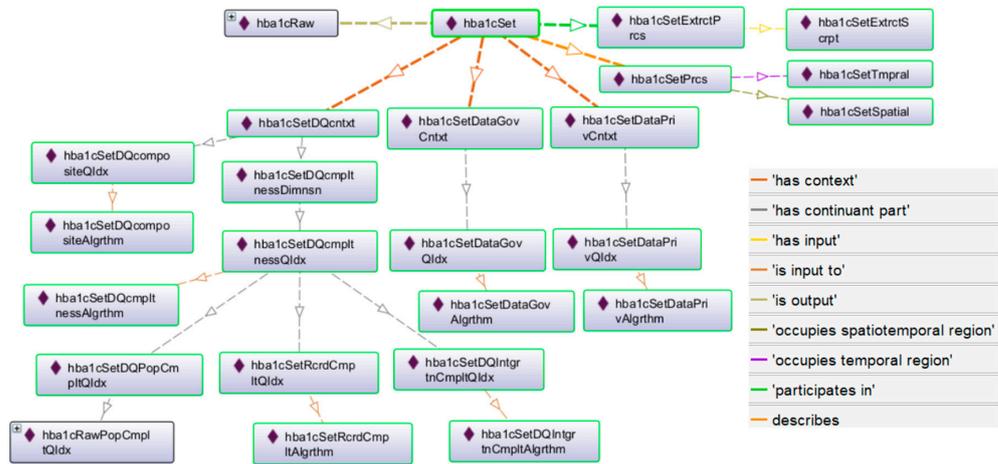
Figure 4. OntoGraf image of the individuals associated with the contextualised dataset "hba1cSet".

A composite quality index might also be required for the whole data quality context (comprising a number of main-level DQ dimensions). This is illustrated in Figure 4 with reference made from the entity "hba1cSetDQcntxt" (relating the to the entire data quality context space) to a composite quality index entity. The data provider therefore has the possibility to set up quality metrics at any stage in the quality-dimension hierarchy and roll them up to composite values at any parent level, while the data user can use the composite values or drill into the individual values to have more granularity.

3.4. Data Governance and Other Dedicated Contexts

A process similar to that followed for the data quality context can be applied to all the other contexts (including the process-related ones) to build up a comprehensive contextual description of the data. Both the data governance and privacy context draw in our example from the methodology described by the privacy and ethics impact and performance assessment (PEIPA) [45]. PEIPA is an interesting example of how individual dimensions can be rolled up to a single metric. It contains 11 main dimensions that are broken into several subdimensions and scored both at the subdimension level and the main dimension level. Most of the dimensions could be made sub-dimensions of the data governance context (e.g., accountability, collection and use, data handling, project approval process) and the privacy context (e.g., consent, safeguarding, anonymisation) and scored according to the PEIPA methodology. This would mean that the sub-dimensions of PEIPA would become a second level of sub-dimensions under the context spaces.

The data governance context could also reference the federated data processes described in [7] relating to the two-stage federated data quality assessment (FedDQA) method. The metrics suggested in this method for scoring completeness, repeatability, accuracy, consistency, degree of structuring, and timeliness could then be referenced by the associated data quality context in the manner described in Section 3.2. Furthermore, the quality improvement processes of other DQ frameworks could form part of the data governance context and be referenced from it.

In addition, quality contexts can as easily well be applied to components of the general context fields. For example, one of the latter concerns the entity geospatial region. A context space could be referenced from this entity via a "has context" relationship and point to the ontology-based spatial data quality assessment framework of Yilmaz et al. [46].

3.5. Retrieving the Contextual Information

Once the data has been contextualised in this manner, all the related information is retrievable and since it is defined in terms of internationalized resource identifiers (IRIs), it can be dereferenced as a whole or on a need-to-know basis.

Dereferencing the information is a relatively straightforward task that can be performed by drilling into the contextual information via description logic (DL) or SPARQL queries. Table 8 shows the results of a simple SPARQL query (1) to find the context spaces referenced by our example. The prefixes of the table entries refer to namespaces.

$$\text{SELECT ?Pred ?Obj WHERE \{do:hba1cSet ?Pred ?Obj . ?Obj rdf:type slcto:Context\_Space . \}} \tag{1}$$

Table 8. Results of SPARQL query (1) showing the context spaces referenced by the "hba1cSet" entity.

Pred	Obj
slcto:has_context	do:hba1cSetDQcntxt
slcto:has_context	do:hba1cSetDataGovCntxt
slcto:has_context	do:hba1cSetDataPrivCntxt

Each of the context spaces can then be dereferenced further; for example, Table 9 shows the results of the SPARQL query (2) to find the quality dimensions referenced by the data quality context entity "hba1cSetDQcntxt" in Table 8.

$$\text{SELECT ?Pred ?Obj WHERE \{do:hba1cSetDQcntxt ?Pred ?Obj . ?Obj rdf:type slcto:Quality\_Dimension . \}} \tag{2}$$

Table 9. Results of SPARQL query (2) showing the quality dimensions referenced in the data quality entity "hba1cSetDQcntxt". The relation "BFO_0000178" refers to the relation "has_continuant part" (following the principles of the OBO Foundry [47]).

Pred	Obj
obo:BFO_0000178	do:hba1cSetDQcmpltnessDimnsn

The querying process may be performed via a recursive search (using an appropriate OWL API interface such as OWL-API or Owlready2) to return all the contextual information describing a dataset (including all the contextual information contained in the whole data chain). Alternatively, information may be dereferenced on a need-to-know basis by querying the relevant subset of entities as shown in Tables 8 and 9. Moreover, since many of the contextualisation entities are sub-typed from the ISO-IEC 11179 tripartite data element, more information can be gleaned from the respective definitions. This would allow users a comprehensive overview of any contextual element, including definitions, references, permissible data values, etc.

3.6. Distinction Between Data-Application Quality Requirements and Data-Entity Quality

We purposefully decoupled the way in which the underlying quality of a data entity is described from the way in which a data application prescribes its quality needs. We have focused mainly on the former since the requirements of the latter are generally specific to the data application.

However, a FAIR-oriented framework can provide some objective means of determining what are otherwise quite subjective measures. As an example, it is possible to provide some quantitative means of measuring the dimension "Understandability". For instance, a data application would itself define what understandability meant for its particular context and then could award points and weightings for the presence of certain descriptive attributes describing standard relations such as "has description", "has scope", "has examples of use", "has data structure", etc. The query outputs could be parsed automatically to find the attributes found and then scored and weighted accordingly to provide an overall measure. The strength of a framework such as SOLICIT is that it is built on a coherent set of attributes inherited from common standards and building blocks that facilitate the construction of general queries to determine such information. The same process can be followed for the other three dimensions in the first column of Table 4. The dimension "Usefulness" could query the descriptive relations concerning scope, description, and the temporal and spatial extent of the data. The dimension "Efficiency", depending on the needs of the data application and the associated data entity/entities employed, could dereference the chain of processes and timestamps to any desired point and make some quantitative measure based on the processing time. Finally, the dimension "Quantity" could dereference relations associated with data completeness, size of data entity, geolocation, and time periods.

Furthermore, given the extent and linkage of the contextualised information, automated AI tools could be used to digest the data and provide reasoned sets of metrics. The advantage is that the contextualisation entities are structured and formalised in DL providing reasoning engines a more straightforward task in making associations without having to formulate their own contextual grids on the parsed textual information.

Although more restrictive, reasoning could also be performed within the ontology itself based on the axiomatic logic. The highlighted bottom line in Figure 3 shows how the entity "hba1cSet" contextualising our data entity example has been classified by an ontology reasoner (ELK) under the class "InformativenessCheck" to provide some indication that the dataset provides a level of associated information. Figure 5 gives the reasoner's explanation for the classification, with explicit reference to the rule on the penultimate two lines. The rule requires the presence of at least the attributes "represents", "uses", and "describes". Rules such as this can be constructed to express minimum acceptance requirements for data, based on the presence of descriptive attributes. More complex rules can be defined but depending upon the logical operators employed, reasoning speeds could be a limiting factor. The ELK reasoner performs very efficiently but is restricted to a subset of logical operators (conjunction, existential restriction, and complex role inclusions). Greater expressivity would require use of other reasoners (e.g., Fact++, Hermitt, Pellet, etc.), and each has their strengths and limitations [48]. An alternative is to handle reasoning using a computer programme interfacing with the ontology in one of the OWL APIs.

Figure 5. Explanation provided by the ontology reasoner for classifying the entity "hba1cSet" under the minimum requirement for an informativeness quality dimension.

3.7. Application of the Concepts to the Indicator Use Case

Considering the seven points listed for a contextualisation paradigm for indictors described in the Introduction, we have been able to show the ability of a user or an intelligent agent to drill into the context down to the granularity provided by the data provider using standard tools that can be automated. The linkage and descriptions of each metadata element dereferenced can be used to draw further connections and associations. For example, the description of our HbA1c example can reference—via the simple knowledge organization system (SKOS) relations established in the contextualisation parameters described in [39]—terminology servers such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) by pointing to the URI: http://purl.bioontology.org/ontology/SNOMEDCT/43396009 (accessed on 12 August 2025). It could also link to other resources such as the logical observation identifiers names and codes (LOINC) to reference the LOINC code 4548-4 or other LOINC codes for HbA1c measurements with different precision. These standard codes than can be used to match other resources that LLMs or other AI tools could use to make further associations in relation to the quality of a data entity against standard definitions.

We have also shown how it is possible to set up quality contexts on a per-data-entity basis and provide the freedom for the data provider to specify and tailor the quality dimensions and metrics relevant for the data entity. This freedom extends to creating further dimensions where required providing a scalable framework to deal with new data paradigms. The quality context can just as easily be applied to a data process. These contexts can be integrated across a whole data chain as discussed in Section 3.1 to provide composite measures of quality that can be unpacked into the individual measures.

The versatility of the approach would have many uses, particularly in linking related datasets from different domains and providing some qualitative understanding of how the datasets could be integrated based on the contextual descriptions and quality parameters provided.

4. Discussion

The work we have presented here resulted from our difficulty in finding an appropriate DQ framework to apply to the use case described in the Introduction. This prompted our reflection on what sort of DQ framework would work for a chain of data processing stages where data users could have access to the quality metrics of each stage, where the quality dimensions and metrics would likely be described and measured in different ways depending upon the data entity type. Additionally, since DQ requirements depend critically on the purpose of the data application, data users should have more granular access to the underlying factors contributing to a quality label that might either be too strict for some applications or too loose for others.

By taking a holistic view of data, a data contextualisation framework has the inherent means to address many of the limitations described in Section 2 and allow the flexibility to specify data quality in the most appropriate way for any given data entity without being prescriptive. It also decouples the intrinsic quality dimensions of the data from those of the data application, allowing many applications to use the same data and make their own conclusions about the data's fitness-for-use and facilitating secondary-data usage.

In view of the overall lack of data contextualisation tools, we extended the SOLICIT framework to show how such a framework could implement data quality in a pragmatic way. This was not intended as a comprehensive demonstration, but rather to show some of the advantages of addressing data quality in this way. Nor does the approach preclude using elements of existing DQ frameworks since SOLICIT is able to reference external methodologies.

The data contextualisation approach addresses several important principles highlighted in the reviews of DQ frameworks, the following in particular:

1. It has relevance to the particular quality dimensions of the data entity being described. For example, completeness is relevant only for certain data entities; it does not carry any meaning for an atomic data type. Completeness can itself be differentiated by concepts such as population completeness, record completeness, data-integration completeness, etc. Since the contextual information is described for a particular data entity, the semantic links are free to change with the type of data entity and to describe only the quality dimensions that are meaningful.

2. It provides the means of describing how a given quality dimension is calculated. SOLICIT provides the functionality to reference a procedure, rule, or algorithm regarding the calculation of any given dimension, also at a composite level. This could be a standard algorithm or a specific one for the needs and could accommodate unstructured data with the associated relevant dimensions and related metrics [49,50].

3. It allows a data entity to associate with any number of context spaces and any number of dimensions within those context spaces. As such, it is not prescriptive but only provides a generic framework that can be applied to the specificities of the type of data entity and data structure. It can therefore support structured and unstructured data, objective and subjective quality metrics, and describing data entity types from an atomic level to dataset level and beyond.

4. It can deal with quality of processes in exactly the same way as it deals with quality of data through the association of quality contexts, quality dimensions, and quality indices.

Wider consensus is however necessary on the nature of the context spaces and on the main-level DQ dimensions they contain. The latter (with some interchanges) were taken from the model of Miller et al. [9] and supplemented with the missing FAIR dimensions. Consensus is also required on the sub-dimensions.

We have nevertheless shown how different dimensions and metrics can be applied to different data entity types within the same framework in a way that is not prescriptive and yet uses a standard set of relationships. We have shown how such a framework could address a chain of data-processing stages and provide provenance of data quality and, most importantly, we have shown a means of integrating data quality with the FAIR data principles in a practical tool.

An important aspect on which we have only touched concerns the possibility of applying learning methods to the contextualised data. As LLMs develop and become more established, the structured means of linking and describing data through standard methodologies would allow such methods quickly to assimilate the entire data ecospace and position the data in an expanded context that could potentially bridge and link domains. Until there are semantic processing languages that can sequentially process natural language inputs and return consistent results [51], the formal contextual constructs of an ontology-based data contextualisation framework will be an important element for LLMs to return consistent results. In this regard, describing the quality of data in a way that would allow comparisons of quality parameters over heterogeneous datasets would be a critical component in ascertaining the suitability of reasoning on combined data.

Limitations and Future Research Directions

We have only skirted the discussion on the exact description of quality dimensions and have not considered any of the possible metrics for measuring them since this was out of our immediate scope. The data contextualisation framework we have introduced is agnostic of such issues and is free to reference any particular methodology and any mappings between them. One of the strengths of the framework we used is that it is extendible at a federated

domain and sub-domain level and can consequently be tailored to domain-specific needs without strict constraint to any prescribed formulation. We hope to address these more pragmatic concerns when contextualising DQ parameters in a real-world indicator as a following step.

We have not considered the real-time data flow processes in the model, but the consolidated data in terms of average measures over a time-period (forming a static dataset) could be contextualised in the manner described in Section 3 and average latency could form one of the associated quality dimensions [5].

We have only described in a precursory way how the DQ requirements of a data user or data application could be specified. Our view is that since these will be quite subjective matters, it would be difficult to quantify them in a rigorous way. Data appropriate for one application may not be appropriate for another. However, the quality indices introduced in the SOLICIT framework provide the means of understanding how a quality score has been determined. The subjective dimensions can be ascertained from the set of general contextual relations describing such aspects as the scope, purpose, definitions of variables, timestamps, geolocation, measurement units, biases, processing steps, etc.

Finally, the work we have undertaken here highlights the general lack of data contextualisation frameworks without which the possibility of realising fully FAIR data is limited. A suitable framework should offer a comprehensive set of standard attributes to contextualise data of widely differing characteristics in a wide variety of domains as touched upon at the beginning of the Introduction. It should not be prescriptive but permit the addition of new dimensions as new data paradigms become available without any major refactoring work. It should also allow for the description of data processes and references to original data from which the data have been derived. Whereas SOLICIT goes someway to fulfilling these requirements, it is not an easy framework to use and requires knowledge of concepts that are not straightforward to assimilate especially in the integration of CCO and ISO/IEC 11179. This could be aided by the development of a graphical user-interface to present the ontology classes in a simplified and more intuitive way as well as to avoid integrity issues when editing the ontology directly. The implementation of SOLICIT in an operational environment would require development and maintenance of the ontology extensions at the domain level. In this regard, SOLICIT is no different to CCO; however, it would require some familiarity with the concepts of the ISO/IEC 11179 metadata registry standard. Nevertheless, in view of the lack of other tools, SOLICIT does furnish a general-purpose data contextualisation framework that can be extended at any domain level allowing the standardisation of quality aspects according to a philosophy grounded on the four foundational principles of FAIR data.

5. Conclusions

The growing importance of interconnected data requires a means of adequately describing the veracity of the underlying data components. Although instrumental in their specific points of focus, existing DQ frameworks are generally context and data-entity dependent. They also tend to disagree on the quality dimensions that should be included.

The lack of convergence to a single DQ framework motivates the quest for an integrated approach. We have attempted to show from the foregoing arguments that many of the contentions between different DQ frameworks can be resolved by decoupling the associated dependencies and dealing with them at an appropriate level of abstraction. A critical step in this regard is to switch the focus from a model that views everything about data as a quality issue to a more holistic data-centric approach that addresses the chief motivation of data reusability from the point of view of the FAIR data principles and

allows the quality to be assessed from comprehensive contextual information provided as metadata.

A framework developed along these lines allows a separation between data-application requirements and a description of the actual data, which avoids the need to describe the quality of data according to different frameworks dependent on the type of data and the type of data application. It also results in a simpler model since most of the quality dimensions considered in DQ frameworks can be resolved into general data contextualisation components.

In view of the general lack of data contextualisation frameworks, we hope the ideas presented here will engender further debate and research on an issue that has far-reaching implications, especially in the linkage of datasets across different domains.

Author Contributions: Conceptualization, N.N. and I.Š.; methodology, N.N., R.N.C. and I.Š.; software, N.N.; validation, R.N.C.; formal analysis, N.N.; data curation, N.N.; writing—original draft preparation, N.N., R.N.C., and I.Š.; writing—review and editing, N.N., R.N.C. and I.Š. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in JRC data browser at https://data.jrc.ec.europa.eu/collection/id-00410 (accessed on 12 August 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application programming interface
BFO	Basic formal ontology
CCO	Common core ontologies
CDE	Common data element
DQ	Data-quality
DL	Description logic
FAIR	Findable, accessible, interoperable, reusable
IRI	Internationalized resource identifier
LLM	Large language model
LOINC	Logical Observation Identifiers Names and Codes
NCD	Non-communicable disease
OWL	Web ontology language
PEIPA	Privacy and ethics impact and performance assessment
SKOS	Simple knowledge organization system
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SOLICIT	Semantic ontology-labelled indicator contextualisation integrative taxonomy
SPARQL	SPARQL query language
URI	Uniform resource identifier
URL	Uniform resource locator

References

1. DAMA UK. The Six Primary Dimensions for Data Quality Assessment—Defining Data Quality Dimensions. Available online: https://www.dama-uk.org/resources/the-six-primary-dimensions-for-data-quality-assessment (accessed on 6 June 2025).
2. *ISO/IEC 25012:2008*; Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)—Data Quality Model. International Organization for Standardization: Geneva, Switzerland, 2008.
3. Byabazaire, J.; O'Hare, G.M.P.; Collier, R.; Delaney, D. IoT Data Quality Assessment Framework Using Adaptive Weighted Estimation Fusion. *Sensors* **2023**, *23*, 5993. [CrossRef]
4. Zhang, L.; Jeong, D.; Lee, S. Data Quality Management in the Internet of Things. *Sensors* **2021**, *21*, 5834. [CrossRef]
5. Jánki, Z.R.; Bilicki, V. A Data Quality Measurement Framework Using Distribution-Based Modeling and Simulation in Real-Time Telemedicine Systems. *Appl. Sci.* **2023**, *13*, 7548. [CrossRef]
6. Zou, K.H.; Berger, M.L. Real-World Data and Real-World Evidence in Healthcare in the United States and Europe Union. *Bioengineering* **2024**, *11*, 784. [CrossRef]
7. Shen, J.; Zhou, S.; Xiao, F. Research on Data Quality Governance for Federated Cooperation Scenarios. *Electronics* **2024**, *13*, 3606. [CrossRef]
8. Haug, A. Understanding the differences across data quality classifications: A literature review and guidelines for future research. *Ind. Manag. Data Syst.* **2021**, *121*, 2651–2671. [CrossRef]
9. Miller, R.; Whelan, H.; Chrubasik, M.; Whittaker, D.; Duncan, P.; Gregório, J. A Framework for Current and New Data Quality Dimensions: An Overview. *Data* **2024**, *9*, 151. [CrossRef]
10. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
11. GO FAIR. R1.3: (Meta) Data Meet Domain-Relevant Community Standards. Available online: https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/ (accessed on 6 June 2025).
12. Cichy, C.; Rass, S. An Overview of Data Quality Frameworks. *IEEE Access* **2019**, *7*, 24634–24648. [CrossRef]
13. Bian, J.; Lyu, T.; Loiacono, A.; Viramontes, T.M.; Lipori, G.; Guo, Y.; Wu, Y.; Prosperi, M.; George, T.J.; Harle, C.A.; et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J. Am. Med. Inf. Assoc.* **2020**, *27*, 1999–2010. [CrossRef]
14. Declerck, J.; Kalra, D.; Vander Stichele, R.; Coorevits, P. Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews. *JMIR Med. Inform.* **2024**, *12*, e51560. [CrossRef]
15. Miller, R.; Chan, S.H.M.; Whelan, H.; Gregório, J. A Comparison of Data Quality Frameworks: A Review. *Big Data Cogn. Comput.* **2025**, *9*, 93. [CrossRef]
16. Lee, Y.W.; Diane, M.; Strong, D.M.; Kahn, B.K.; Wang, R.Y. AIMQ: A methodology for information quality assessment. *Inf. Manag.* **2002**, *40*, 133–146. [CrossRef]
17. Batini, C.; Cabitza, F.; Cappiello, C.; Francalanci, C. A comprehensive data quality methodology for web and structured data. *Int. J. Innov. Comput. Appl.* **2008**, *1*, 205–218. [CrossRef]
18. Loshin, D. Economic framework of data quality and the value proposition. In *Enterprise Knowledge Management: The Data Quality Approach*; Morgan Kaufmann Publisher: San Francisco, CA, USA, 2001; pp. 73–99. [CrossRef]
19. Pipino, L.L.; Lee, Y.W.; Wang, R.Y. Data quality assessment. *Commun. ACM* **2002**, *45*, 211–218. [CrossRef]
20. Sebastian-Coleman, L. *Measuring Data Quality for Ongoing Improvement*; Morgan Kaufmann Publisher: Waltham, MA, USA, 2013.
21. Del Pilar Angeles, M.; García-Ugalde, F. A data quality practical approach. *Int. J. Adv. Softw.* **2009**, *2*, 259–274.
22. Batini, C.; Barone, D.; Cabitza, F.; Grega, S. A data quality methodology for heterogeneous data. *Int. J. Database Manage. Syst.* **2011**, *3*, 60–79.
23. Cappiello, C.; Ficiaro, P.; Pernici, B. HIQM: A methodology for information quality monitoring, measurement, and improvement. In *Advances in Conceptual Modeling-Theory and Practice: ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT, Tucson, AZ, USA, 6–9 November 2006*; Proceedings 25; Springer Publisher: Berlin/Heidelberg, Germany, 2006; pp. 339–351.
24. Sundararaman, A.; Venkatesan, S.K. Data quality improvement through OODA methodology. In Proceedings of the 22nd MIT International Conference on Information Quality, ICIQ, Rock, AR, USA, 6–7 October 2017; pp. 1–14.
25. Vaziri, R.; Mohsenzadeh, M.; Habibi, J. TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement. *PLoS ONE* **2016**, *11*, e0154508. [CrossRef]
26. Wang, R.Y. A product perspective on total data quality management. *Commun. ACM* **1998**, *41*, 58–65. [CrossRef]
27. English, L.P. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*; John Wiley & Sons, Inc. Publisher: New York, NY, USA, 1999.
28. *ISO 8000-8:2015*; Data Quality—Part 8: Information and Data Quality: Concepts and Measuring. ISO Publisher: Geneva, Switzerland, 2015.

29. Federal Privacy Council. Fair Information Practice Principles (FIPPS). 2024. Available online: https://www.fpc.gov/resources/fipps/ (accessed on 6 June 2025).

30. European Commission. Quality Assurance Framework of the European Statistical System v2.0. 2019. Available online: https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf (accessed on 6 June 2025).

31. Government Data Quality Hub. The Government Data Quality Framework. 2020. Available online: https://www.gov.uk/government/publications/the-government-data-quality-framework (accessed on 6 June 2025).

32. DAMA International. *DAMA-DMBOK Data Management Body of Knowledge*, 2nd ed.; Technics Publications: Sedona, AZ, USA, 2017. Available online: https://technicspub.com/dmbok/ (accessed on 6 June 2025).

33. International Monetary Fund. Data Quality Assessment Framework (DQAF). 2003. Available online: https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm (accessed on 6 June 2025).

34. Basel Committee on Banking Supervision. *Principles for Effective Risk Data Aggregation and Risk Reporting*; Technical Report; Bank for International Settlements: Basel, Switzerland, 2013. Available online: https://www.bis.org/publ/bcbs239.htm (accessed on 6 June 2025).

35. Choudhary, A. ALCOA and ALCOA Plus Principles for Data Integrity. 2024. Available online: https://www.pharmaguideline.com/2018/12/alcoa-to-alcoa-plus-for-data-integrity.html (accessed on 6 June 2025).

36. World Health Organization. *Data Quality Assurance: Module 1: Framework and Metrics*; World Health Organization: Geneva, Switzerland, 2022; 30p. Available online: https://www.who.int/publications/i/item/9789240047365 (accessed on 6 June 2025).

37. Karr, A.F.; Sanil, A.P.; Banks, D.L. Data quality: A statistical perspective. *Stat. Methodol.* **2006**, *3*, 137–173. [CrossRef]

38. Chen, H.; Hailey, D.; Wang, N.; Yu, P. A review of data quality assessment methods for public health information systems. *Int. J. Environ. Res. Public Health* **2014**, *11*, 5170–5207. [CrossRef]

39. Nicholson, N.; Štotl, I. A generic framework for the semantic contextualization of indicators. *Front. Comput. Sci.* **2024**, *6*, 1463989. [CrossRef]

40. ISO/IEC 11179:2015; Information Technology—Metadata Registries (MDR)—Part 1: Framework. Available online: https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-1:ed-3:v1:en (accessed on 6 June 2025).

41. Jensen, M.G.; De Colle Kindya, S.; More, C.; Cox, A.P.; Beverley, J. The common core ontologies. *arXiv* **2024**. [CrossRef]

42. CUBRC, Inc. An Overview of the Common Core Ontologies. 2019. Available online: https://www.nist.gov/system/files/documents/2021/10/14/nist-ai-rfi-cubrc_inc_004.pdf (accessed on 31 July 2025).

43. Corrales, D.C.; Ledezma, A.; Corrales, J.C. From Theory to Practice: A Data Quality Framework for Classification Tasks. *Symmetry* **2018**, *10*, 248. [CrossRef]

44. Sacks, D.B.; Arnold, M.; Bakris, G.L.; Bruns, D.E.; Horvath, A.R.; Lernmark, A.; Metzger, B.E.; Nathan, D.M.; Kirkman, M.S. Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. *Diabetes Care* **2023**, *46*, e151–e199. [CrossRef]

45. Di Iorio, C.T.; Carinci, F.; Oderkirk, J.; Smith, D.; Siano, M.; de Marco, D.A.; de Lusignan, S.; Hamalainen, P.; Benedetti, M.M. Assessing data protection and governance in health information systems: A novel methodology of Privacy and Ethics Impact and Performance Assessment (PEIPA). *J. Med. Ethics* **2021**, *47*, e23. [CrossRef] [PubMed]

46. Yılmaz, C.; Cömert, Ç.; Yıldırım, D. Ontology-Based Spatial Data Quality Assessment Framework. *Appl. Sci.* **2024**, *14*, 10045. [CrossRef]

47. Open Biological and Biomedical Ontology Foundry. Principles: Overview. Available online: http://obofoundry.org/principles/fp-000-summary.html (accessed on 1 August 2025).

48. Dentler, K.; Cornet, R.; ten Teije, A.; de Keizer, N. Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. *Semant. Web* **2011**, *2*, 71–87. Available online: https://www.semantic-web-journal.net/sites/default/files/swj120_2.pdf (accessed on 12 August 2025). [CrossRef]

49. Ramasamy, A.; Chowdhury, S. Big Data Quality Dimensions: A Systematic Literature Review. *J. Inf. Syst. Technol. Manag.* **2020**, *17*, e202017003. [CrossRef]

50. Elouataoui, W.; El Alaoui, I.; El Mendili, S.; Gahi, Y. An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data Cogn. Comput.* **2022**, *6*, 153. [CrossRef]

51. Talburt, J. Data Speaks for Itself: Data Quality Management in the Age of Language Models. The Data Administration Newsletter. Available online: https://tdan.com/data-speaks-for-itself-data-quality-management-in-the-age-of-language-models/32410 (accessed on 12 August 2025).