Data Sci. Eng. (2016) 1(1):6–20
DOI 10.1007/s41019-015-0004-7

INVITED PAPER

On the Meaningfulness of “Big Data Quality” (Invited Paper)

Donatella Firmani[1] · Massimo Mecella[2] · Monica Scannapieco[3] · Carlo Batini[4]

Received: 1 September 2015 / Revised: 2 October 2015 / Accepted: 2 October 2015 / Published online: 5 November 2015

Abstract In this paper, we discuss the application of concept of data quality to big data by highlighting how much complex is to define it in a general way. Already data quality is a multidimensional concept, difficult to characterize in precise definitions even in the case of well-structured data. Big data add two further dimensions of complexity: (i) being “*very*” *source specific*, and for this we adopt the interesting UNECE classification, and (ii) being *highly unstructured and schema-less*, often without golden standards to refer to or very difficult to access. After providing a tutorial on data quality in traditional contexts, we analyze big data by providing insights into the UNECE classification, and then, for each type of data source, we choose a specific instance of such a type (notably deep Web data, sensor-generated data, and Twitters/short texts) and discuss how quality dimensions can be defined in these cases. The overall aim of the paper is therefore to identify further research directions in the area of big data quality, by providing at the same time an up-to-date state of the art on data quality.

Keywords Data quality · Big data · Quality dimensions · Information quality

1 Introduction

We are currently in the era of big data [21], and whereas no exact definition of what big data are has been agreed upon in the research community and among practitioners, the “common sense” suggests that large data sets, featuring 3 V’s (volume, variety, and velocity), and interesting for analytic tasks that can be carried out over them in order to discover interesting patterns, are what defines big data.

The common vision of big data emphasizes quantity over *quality* of data, arguing that the very large amount of data is sufficient to offset any distortion or defects that data might contain. This view is probably too simplistic, and critical research direction is to develop effective and efficient methods for assessing the quality of data and the reliability of inferences made through quality-aware algorithms.

But what is then *big data quality*? In this paper, we will show, in a nonformal way but trough examples and case studies, how difficult is to define a unique concept of big data quality. We will argue that devising a unique data quality concept is not meaningful at all. Conversely, there are many notions of quality, to be applied to specific types of big data, that should be carefully considered when dealing with big data sets and analytics over them.

The rest of this paper is structured as it follows: Sect. 2 introduces the notion of data and information quality, as addressed in general terms by the literature, in order to pro-

D. Firmani: Partially supported by MIUR under project AMANDA (Algorithmics for MAssive and Networked DAta).
M. Mecella: Partially supported by the EU project VOICE (Grant Agreement 621137) and the Italian projects NEPTIS and RoMA and the cluster Social Museum and Smart Tourism.
C. Batini: Partially supported by the EU project COMSODE (Grant Agreement 611358).

✉ Massimo Mecella
mecella@dis.uniroma1.it

Donatella Firmani
firmani@ing.uniroma2.it

Monica Scannapieco
scannapi@istat.it

Carlo Batini
batini@disco.unimib.it

[1] Università di Roma Tor Vergata, Rome, Italy

[2] Sapienza Università di Roma, Rome, Italy

[3] Istituto Nazionale di Statistica (ISTAT), Rome, Italy

[4] Università di Milano Bicocca, Milan, Italy

Table 1 A relation Movies with data quality problems

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	**Weir**	**1942**	3	**1940**
2	Dead poets society	**Curtiz**	1989	0	Null
3	**Rman Holiday**	Wylder	1953	0	Null
4	Sabrina	**Null**	1964	**0**	**1985**

vide a basic background; then, Sect. 3 focuses on big data, by referring to the UNECE classification, in order to provide a discussion of big data quality in the Sect. 4. Section 5 concludes the paper by highlighting its main contributions and identifying further research directions for quality of big data.

2 Data and Information Quality

Quality, in general, has been defined as the "totality of characteristics of a product that bear on its ability to satisfy stated or implied needs" [29], as "fitness for (intended) use" [31], "conformance to requirements" [9], "user satisfaction" [52].

When people think about *information quality*, they often reduce quality just to *accuracy*, e.g., the city name "Chicago" misspelled as "Chcago." Indeed, information is normally considered to be of poor quality if typos are present or wrong values are associated with a concept instance, such as an erroneous birth date or age associated with a person. Hence, information quality is more than simply accuracy. Other significant dimensions such as completeness, consistency, and currency are necessary in order to fully characterize the quality of information. Table 1 provides some examples of these dimensions for structured data, from [1]. The relational table describes movies, with title, director, year of production, number of remakes, and year of the last remake. The cells with data quality problems are bold faced. At first, only the cell corresponding to the title of movie 3 seems to be affected by a data quality problem. Indeed, there is a misspelling in the title, where Rman stands for Roman, thus causing an accuracy problem. Nevertheless, another accuracy problem is related to the exchange of the director between movies 1 and 2: Weir is actually the director of movie 2 and Curtiz the director of movie 1. Other data quality problems are: a missing value for the director of movie 4, causing a completeness problem, and a 0 value for the number of remakes of movie 4, causing a currency problem because a remake of the movie has actually been proposed. Finally, there are two consistency problems: first, for movie 1, the value of LastRemakeYear cannot be lower than Year; second, for movie 4, the value of LastRemakeYear cannot be different from null, because the value of #Remakes is 0.

The above examples and considerations show that:

- Data quality is a multifaceted concept, and different dimensions concur to define it;
- Quality problems related to some dimensions, such as accuracy, can be easily detected in some cases (e.g., misspellings) but are more difficult to detect in other cases (e.g., where admissible but not correct values are provided);
- A simple example of a completeness error has been shown, but as it happens with accuracy, completeness can also be very difficult to evaluate (e.g., if a tuple representing a movie is entirely missing from the relation Movie);
- Consistency detection does not always localize the errors (e.g., for movie 1, the value or the LastRemakeYear attribute is wrong).

The described examples concern a table in a relational database. Problems change significantly when other *types of information*, different from relational data, are involved. An unbelievable vast amount of information about realities of interest is indeed represented by information which is not encoded as structured data. Reality is typically represented by a piece of information either in its realistic inherent character (e.g., a photograph of a landscape or a photograph of a group of students in a class or a map and a descriptive text in a travel guide) or in other ways, e.g., in novels and poetry as a virtual representation of the reality itself. Hence, quality issues and techniques may differ depending on the information representation, e.g., images, maps, and unstructured text.

Therefore, in what follows, we briefly characterize the concept of information quality for both structured data (this is referred specifically as *data quality*) and other types of information, by defining information quality's dimensions, together with possible metrics to measure them. We remark that the purpose of the section is not to provide a complete overview of the information quality concept (the interested reader can refer to the book [1] for a comprehensive technical coverage of information quality), but rather to allow the reader to understand the intrinsic complexity of the concept itself, and the many facets that information quality can have. On the basis of such a characterization, next sections will elaborate on the concept of information quality in a specific way for elected types of big data sources.

2.1 On the Definition and Measurement of Information Quality: Dimensions and Metrics

Dimensions for information quality can be grouped into clusters according to [2]. Dimensions are included in the same cluster according to their similarity with respect to their ability to capture an information quality aspect. Clusters are defined in the following list, where the first item in italics is the representative dimension of the cluster, followed by other member dimensions, namely:

1. *Accuracy*, correctness, validity, and precision focus on the adherence to a given reality of interest.
2. *Completeness*, pertinence, and relevance refer to the capability of representing all and only the relevant aspects of the reality of interest.
3. *Consistency*, cohesion, and coherence refer to the capability of the information to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules, and other formalisms.
4. *Redundancy*, minimality, compactness, and conciseness refer to the capability of representing the aspects of the reality of interest with the minimal use of informative resources.
5. *Readability*, comprehensibility, clarity, and simplicity refer to ease of understanding and fruition of information by users.
6. *Accessibility* and availability are related to the ability of the user to access information from her own culture, physical status/functions, and technologies available.
7. *Trust*, including believability, reliability, and reputation, focuses on how much information derives from an authoritative source.
8. *Usefulness* is related to the advantage the user gains from the use of information.

Dimensions are usually defined in a qualitative way, referring to general properties of data, and the related definitions do not provide any tool or methodology for assigning values to the dimensions themselves. Specifically, definitions do not typically provide quantitative measures, but one or more *metrics* are to be associated with dimensions as separate, distinct properties. In the following, for each cluster of dimensions described above, we provide definitions for some selected dimensions and examples of possible metrics.

Moreover, as shown in Table 2, the first three dimensions are specifically discussed for structured data, the fourth dimension for linked data (as structured Web data), the fifth dimension for texts (unstructured data), the sixth and seventh dimensions for Web data in general, and the eighth dimension for images. The rationale for this choice is to give the reader an overview as richer as possible of what information quality means for different information types, and to consider information types which are relevant in the context of big data sources.

Table 2 Clusters and information types

Cluster	Information type
Accuracy	Structured data
Completeness	Structured data
Consistency	Structured data
Redundancy	Linked data—structured Web data
Readability	Texts—unstructured data
Accessibility	Web sites' data
Trust	Web data sources
Usefulness	Images

2.1.1 The Accuracy Cluster

Accuracy is defined as the closeness between a data value `v` and a data value `v'`, considered as the correct representation of the real-life phenomenon that the data value `v` aims to represent. As an example, if the name of a person is `John`, the value `v'` = `John` is correct, while the value `v` = `Jhn` is incorrect. The world around us changes (velocity is one of the 3 V's of big data), and what we have referred in the above definition as "the real-life phenomenon that the data value `v` aims to represent" reflects such changes. So, there is a particular yet relevant type of data accuracy that refers to the rapidity with which the change in real-world phenomenon is reflected in the update to the data value; we call *temporal accuracy* such type of accuracy, in contrast to *structural accuracy* (or, simply, *accuracy*), that characterizes the accuracy of data as observed in a specific time frame, where the data value can be considered stable and unchanged. In the following, we consider first structural accuracy and later temporal accuracy. Two kinds of (structural) accuracy can be identified, namely a syntactic accuracy and a semantic accuracy.

Syntactic accuracy is the closeness of a value `v` to the elements of the corresponding definition domain D. In syntactic accuracy, we are not interested in comparing `v` with the true value `v'`; rather, we are interested in checking whether `v` belongs to D, whatever it is. So, if `v` = `Jack`, even if `v'` = `John`, `v` is considered syntactically correct, as `Jack` is an admissible value in the domain of persons' names. Syntactic accuracy is measured by means of functions, called *comparison functions*, that evaluate the distance between `v` and the values in D. The edit distance is a simple example of a comparison function, taking into account the minimum number of character insertions, deletions, and replacements to convert a string `s` to a string `s'`. More complex comparison functions exist, e.g., taking into account similar sounds or character transpositions (see [8]).

Semantic accuracy is the closeness of the value `v` to the *true* value `v'`. Let us consider again the relation `Movies` of Table 1. The exchange of directors' names in tuples 1 and 2 is an example of a semantic accuracy error. Indeed, for movie 1, a director named `Curtiz` would be admissible, and thus, it is syntactically correct. Nevertheless, `Curtiz` is not the director of `Casablanca`; therefore, a semantic accuracy error occurs. The above examples clearly show the difference between syntactic and semantic accuracy. Note that, while it is reasonable to measure syntactic accuracy using a distance function, semantic accuracy is measured better with a `<yes, no>` or a `<correct, not correct>` domain. Consequently, semantic accuracy coincides with the concept of *correctness*. In contrast with what happens for syntactic accuracy, in order to measure the semantic accuracy of a value `v`, the corresponding true value has to be known, or, else, it should be possible, by considering additional knowledge, to infer whether that value `v` is or is not the true value. In a general context, a technique for checking semantic accuracy consists of looking for the same data in different data sources and finding the correct data by comparisons. This latter approach also requires the solution of an *object identification problem*, i.e., the problem of understanding whether two tuples refer to the same real-world entity or not [14].

As anticipated, a relevant aspect of data is their change and update during time. *Temporal accuracy* can be characterized in terms of currency, volatility, and timeliness:

> Currency concerns how promptly data are updated with respect to changes occurred in the real world. As an example in Table 1, the attribute `#Remakes` of movie 4 has low currency because a remake of movie 4 has been done, but this information did not result in an increased value for the number of remakes. Similarly, if the residential address of a person is updated, i.e., it corresponds to the address where the person lives, then the currency is high.
> Volatility characterizes the frequency with which data vary in time. For instance, stable data such as birth dates have volatility equal to 0, as they do not vary at all. Conversely, stock quotes, a kind of frequently changing data, have a high degree of volatility due to the fact that they remain valid for very short time intervals.
> Timeliness expresses how data are *current* for the task at hand. The timeliness dimension is motivated by the fact that it is possible to have current data that are actually useless because they are *late* for a specific usage. For instance, the timetable for university courses is current if contains the most recent data, but it is not timely if it is available only after the start of the classes.

Currency can be typically measured with respect to metadata concerning the *last update*, i.e., the last time stamp at which the specific data were updated. For data types that change with a fixed frequency, the last update metadata allow us to compute currency straightforwardly. Conversely, for data types whose change frequency can vary, one possibility is to calculate an average change frequency and perform the currency computation with respect to it, thus tolerating some errors. As an example, if a data source stores product names that are estimated to change every five years, then a product, having its last update metadata reporting a date corresponding to a month before the observation time, can be assumed to be *current*; conversely, if the date reported is ten years before the observation time, it can be assumed to be *not current*.

Volatility is a dimension that inherently characterizes certain types of data. A metric for volatility is given by the timespan (or its inverse) that data remain valid.

Timeliness implies that data not only are current, but are also in time for events corresponding to their usage. Therefore, a possible measurement consists of (i) a currency measure and (ii) a check that data are available *before* the planned usage time.

2.1.2 The Completeness Cluster

Completeness can be generically defined as "the extent to which data are of sufficient breadth, depth, and scope for the task at hand" [51]. In [43], three types of completeness are identified. *Schema completeness* is defined as the degree to which concepts and their properties are not missing from the schema. *Column completeness* is defined as a measure of the missing values for a specific property or column in a table. *Population completeness* evaluates missing values with respect to a reference population.

If focusing on a specific data model, a more precise characterization of completeness can be given. In the following, we refer to the relational model and to the case of the Closed World Assumption with null values (see [1] for further details). Another example of completeness characterization is related to Web data [42].

In the model with null values with CWA, specific definitions for completeness can be provided by considering the granularity of the model elements, i.e., values, tuples, attributes, and relations, as shown in Fig. 1. Specifically, it is possible to define

- a *value completeness* to capture the presence of null values for some fields of a tuple;
- a *tuple completeness* to characterize the completeness of a tuple with respect to the values of all its fields;
- an *attribute completeness* to measure the number of null values of a specific attribute in a relation;
- a *relation completeness* to capture the presence of null values in a whole relation.

As an example, in Table 3, a `Student` relation is shown. The tuple completeness evaluates the percentage of specified values in the tuple with respect to the total number of attributes of the tuple itself. Therefore, in the example, the tuple completeness is 1 for tuples 6754 and 8907, 0.8 for tuple 6587, 0.6 for tuple 0987, and so on. A possible way to measure the tuple completeness is to measure the information content of the tuple with respect to its maximum potential

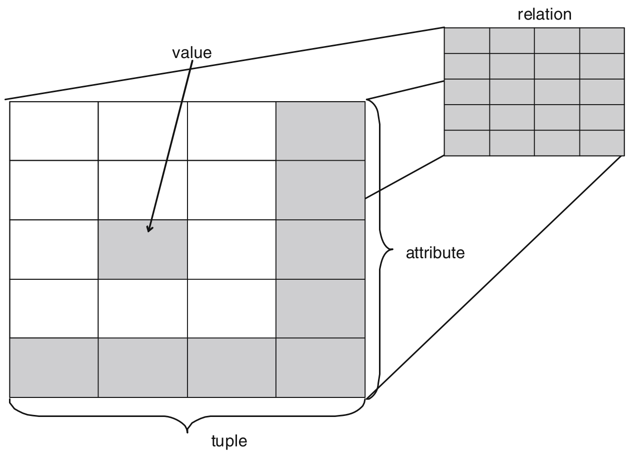


Fig. 1 Completeness of different elements in the relational model

Table 3 `Student` relation exemplifying the completeness of tuples, attributes, and relations

Student ID	Name	Surname	Vote	Examination date
6754	Mike	Collins	29	07/17/2004
8907	Anne	Herbert	18	07/17/2004
6578	Julianne	Merrals	Null	07/17/2004
0987	Robert	Archer	Null	Null
1243	Mark	Taylor	26	09/30/2004
2134	Bridget	Abbott	30	09/30/2004
6784	John	Miller	30	Null
0098	Carl	Adams	25	09/30/2004
1111	John	Smith	28	09/30/2004
2564	Edward	Monroe	Null	Null
8976	Anthony	White	21	Null
8973	Marianne	Collins	30	10/15/2004

information content. With reference to this interpretation, we are implicitly assuming that all values of the tuple contribute equally to the total information content of the tuple. Of course, this may not be the case, as different applications can weight the attributes of a tuple differently.

The attribute completeness evaluates the percentage of specified values in the column corresponding to the attribute with respect to the total number of values that should have been specified. In Table 3, let us consider an application calculating the average of the votes obtained by students. The absence of some values for the `Vote` attribute simply implies a deviation in the calculation of the average; therefore, a characterization of `Vote` completeness may be useful.

The relation completeness is relevant in all applications that need to evaluate the completeness of a whole relation and can admit the presence of null values on some attributes. Relation completeness measures how much information is represented in the relation by evaluating the content of the information actually available with respect to the maximum possible content, i.e., without null values. According to this interpretation, completeness of the relation `Student` in Table 3 is 53/60.

2.1.3 The Consistency Cluster

The consistency captures the violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file. With reference to the relational theory, *integrity constraints* are an instantiation of such semantic rules. The reader can consider [15] for details about consistency detection and correction in the relational model.

In the area of Official Statistics, *data edits* are another example of semantic rules that allow for the checking of consistency. As an example, data coming from survey questionnaires have a structure corresponding to the *questionnaire schema*. The semantic rules are thus defined over such a structure in a way very similar to relational constraints and are called *edits*. *Data editing* is defined as the task of detecting inconsistencies by formulating rules that must be respected by every correct set of answers. Such rules are expressed as *edits*, which denote error conditions. After the detection of erroneous records, the act of correcting erroneous fields by restoring correct values is called *imputation*. The problem of localizing errors by means of edits and imputing erroneous fields is known as the *edit-imputation problem* and solutions to that date back to mid-1970s [17].

2.1.4 The Redundancy Cluster

We define conciseness for the particular case of linked data [5] that have the property to be structured data on the Web. Conciseness for linked data refers to the presence of irrelevant elements with respect to the domain or the minimization of redundant schema and data elements.

There are two major notions of conciseness:

- *intensional conciseness*, which refers to the case when the data set does not contain redundant schema elements (properties and classes). Only essential properties and classes are included in the schema;
- *extensional conciseness*, which refers to the case when the data set does not contain redundant objects (instances).

Intensional conciseness measures the number of unique schema elements (i.e., properties and classes) of a data set in relation to the overall number of schema elements in a schema [39]. Extensional conciseness measures the number of unique entities in relation to the overall number of entities in the data set [39]. Further, extensional conciseness can be measured as the total number of instances that violate the uniqueness rule in relation to the total number of relevant instances [20,35]. An example of intensional conciseness would be a particular flight, e.g., A123, being represented by two different properties in the same data set, such as http://flights.org/airlineID and http://flights.org/name. In this case, redundancy between airlineID and name can ideally be solved by merging the two properties and keeping only one unique identifier. In other words, conciseness should push stakeholders to reuse as much as possible schema elements from existing schemata/ontologies rather than creating new ones since the reuse will support data interoperability.

Representational conciseness refers to the extent to which information is compactly represented. As an example, consider a flight portal that represents the URIs for the destinations compactly with the use of the airport codes, e.g., MXP is the airport code for Milano Malpensa; therefore, the URI is http://airlines.org/MXP. This short representation of URIs helps users share and remember them easily.

Representational conciseness can be measured as: (a) detection of long URIs or those that contain query parameters [26], or (b) detection of RDF primitives, i.e., RDF reification, RDF containers, and RDF collections [26]. The concise representation of data not only contributes to the human readability of that data, but also influences the performance of data when queried. Keeping URIs concise and human readable is highly recommended for large scale and/or frequent processing of RDF data as well as for efficient indexing and serialization.

2.1.5 The Readability Cluster

Readability is a relevant dimension especially for texts; therefore, in the following, the specific focus is on this type of information that is highly unstructured. Readability is defined as reading easiness. Readability is also defined as what makes some texts easier to read than others [13]. [34] defines readability as “the ease of understanding due to the style of writing.” This definition focuses on writing style as separate from issues such as content, coherence, and organization. Readability is then concerned with the relative difficulty of reading written text. Readability should not to be confused with *legibility*, which is concerned with typeface and layout.

Readability research largely traces its origins to an initial study by Kitson [33] that demonstrates tangible differences in sentence lengths and word lengths, measured in syllables, between two newspapers and two magazines (see also [55] for an historical perspective of readability). The majority of metrics proposed for readability are based on factors that represent two broad aspects of comprehension difficulty: (i) lexical or semantic features and (ii) sentence or syntactic complexity. According to [7], formulas that depend on these variables are popular because they are easily associated with text simplification.

As a consequence of the above perspective, readability is usually measured by using a mathematical formula that considers syntactic features of a given text, such as word length and sentence length. Over 200 formulas have been reported for readability in the English language [13] from 1920s to 1980s, among them the Gunning-Fox index [24], the Automated Readability Index (ARI) [46], the Flesch Reading Ease [16,19], and the Flesch Kincaid Grade Level [32]; we briefly discuss a few of them in order to provide some intuition of the ideas behind. The Gunning-Fox index produces an approximate grade level required to understand the document. The basic idea in the index is that the longer sentences are and the greater is the complexity of words used in them, the higher is the difficulty to read the text. The formula for the Gunning-Fox index is shown in Fig. 2.

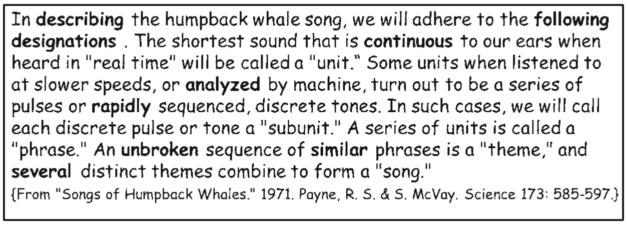
An example of evaluation of the Gunning-Fox index from [41] is the text in Fig. 3. This passage has seven sentences and 96 words. The average sentence length is 13.7. There are nine difficult words (in **boldface**). The Gunning’s Fox index is $= 0.4 \times (13.7 + 9.375) = 9.23$.

ARI is a readability measure designed to represent the US grade level needed to comprehend the text. Unlike the other indexes, ARI relies on a ratio characters per word, instead of the usual syllables per word. See the formula for the ARI index in Fig. 4, where:

– characters are the number of characters in the text;
– words are the number of words in the text;
– sentences are the number of sentences in the text;
– complex words are difficult words defined as those with three or more syllables.

$$\boxed{0.4 * \left[\left(\frac{words}{sentence}\right) + 100 * \left(\frac{complexwords}{words}\right)\right]}$$

Fig. 2 Formula for the Gunning-Fox index



In **describing** the humpback whale song, we will adhere to the **following designations** . The shortest sound that is **continuous** to our ears when heard in "real time" will be called a "unit." Some units when listened to at slower speeds, or **analyzed** by machine, turn out to be a series of pulses or **rapidly** sequenced, discrete tones. In such cases, we will call each discrete pulse or tone a "subunit." A series of units is called a "phrase." An **unbroken** sequence of **similar** phrases is a "theme," and **several** distinct themes combine to form a "song."
{From "Songs of Humpback Whales." 1971. Payne, R. S. & S. McVay. Science 173: 585-597.}

Fig. 3 Example of evaluation of the Gunning-Fox index

$$\boxed{\mathrm{ARI} = 4{,}71 * \frac{\text{characters}}{\text{words}} + 0{,}5 * \frac{\text{complex words}}{\text{sentences}} - 21{,}43}$$

Fig. 4 Formula for the ARI index

2.1.6 The Accessibility Cluster

Publishing large amounts of data in Web sites is not a sufficient condition for their availability to everyone. In order to access it, a user needs to access a network, to understand the language to be used for navigating and querying the Web, and to perceive with senses the information made available. Accessibility measures the ability of the user to access the data from her own culture, physical status/functions, and technologies available. We focus in the following on causes that can reduce physical or sensorial abilities, and, consequently, can reduce accessibility, and we briefly outline corresponding guidelines to achieve accessibility. Among others, the World Wide Web Consortium [50] defines the individuals with disabilities as subjects that (i) may not be able to see, hear, move, or process some types of information easily or at all; (ii) may have difficulty reading or comprehending text; (iii) may not have to or be able to use a keyboard or mouse; (iv) may have a text-only screen, a small screen, or a slow Internet connection; and (v) may not speak or understand a natural language fluently.

Several guidelines are provided by international and national bodies to govern the production of data, applications, services, and Web sites in order to guarantee accessibility. One of the most well-known guidelines related to data accessibility is provided by the World Wide Web Consortium [50]; we will not discuss further as it is out of the scope of this work. Several countries have enacted specific laws to enforce accessibility in public and private Web sites and applications used by citizens and employees in order to provide them effective access and reduce the digital divide.

2.1.7 The Trust Cluster

Trust is a level of subjective and local probability with which an agent assesses that another agent will perform a particular action. *Trustworthiness* is the objective probability that the trustee performs a particular action on which the interests of the truster depend. Though trust and trustworthiness are two distinct concepts, when dealing with techniques for assessing them, the two concepts play often a single role; hence, in the following, the two terms will be used interchangeability unless specific characterizations are needed.

In the following, we elaborate on the three dimensions useful to characterizing trustworthiness, namely *believability*, *verifiability*, and *reputation*.

Believability refers to the extent to which information is regarded as true and credible. Believability can also be defined as the subjective measure of user belief that the data are “true” [30]. An easy way for measuring believability is by checking whether the contributor is contained in a list of trusted providers.

Verifiability refers to the degree by which a data consumer can assess the correctness of a data set. Verifiability is described as the “degree and ease with which the information can be checked for correctness” [4]. Similarly, in [18], the verifiability criterion is used as the means a consumer is provided with, which can be used to examine the data for correctness. Verifiability can be measured either by an unbiased third party, if the data set itself points to the source, or by the presence of a digital signature. A mean for verifying in linked data is to provide basic provenance information along with the data set, such as using existing vocabularies such as SIOC, Dublin Core, Provenance Vocabulary, the OPMV,[1] or the recently introduced PROV vocabulary.[2] Yet another mechanism is the usage of digital signatures [6], whereby a source can sign either a document containing an RDF serialization or an RDF graph.

Reputation is a judgment made by a user to determine the integrity of a source. It can be associated with a data publisher, a person, organization, group of people, or community of practice, or it can be a characteristic of a data set. [22] estimates the reputation of an entity (i.e., a publisher or a data set) either as a result from direct experience or as recommendations from others. They propose the tracking of reputation through a centralized authority or, in alternative, via decentralized voting. There are different possibilities for determining reputation and can be classified into human-based or (semi-) automated approaches. The human-based approach is via a survey in a community or by questioning other members who can help to determine the reputation of a source or by the person who published a data set; conversely, the (semi-) automated approach can be performed by the use of external links or page ranks.

2.1.8 The Usefulness Cluster

We characterize the usefulness by specifically focusing on images. A well-known model for image quality is the Fidelity-Usefulness-Naturalness (FUN, [11]) that assumes the existence of three major dimensions: fidelity, usefulness, and naturalness. Fidelity is the degree of apparent match of the image with the original. Naturalness is the degree of apparent match of the image with the viewer’s internal references. This attribute plays a fundamental role when we have to evaluate the quality of an image without having access to the corresponding original. We provide in the following more details on usefulness. Usefulness is the degree of apparent suitability of the image with respect to a specific task. In many application domains, such as medical or astronomical imaging, image processing procedures can be applied to increase the image usefulness [23]. An example of image usefulness is

[1] http://open-biomed.sourceforge.net/opmv/ns.html.

[2] http://www.w3.org/TR/prov-o/.

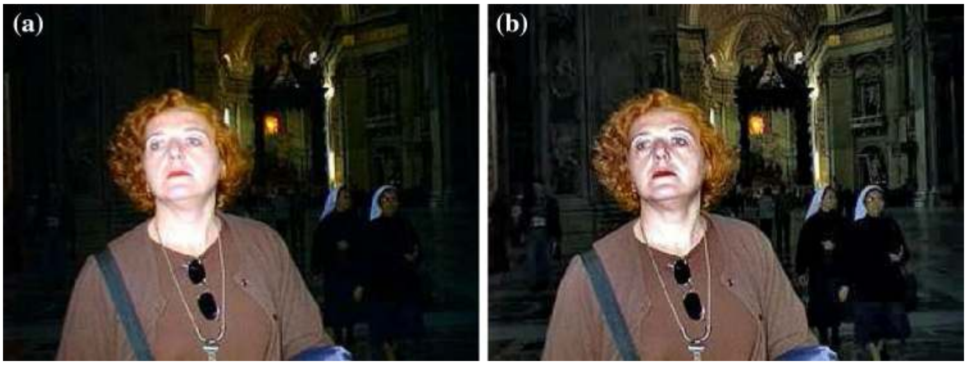


Fig. 5 Example of image usefulness, taken from [1], **a** a faithful image, **b** a contrast-enhanced image showing more details in the background

shown in Fig. 5. The image to the left may be accurate with respect to the original, but the image to the right shows more details in the background due to a contrast enhancement algorithm applied. The enhancement processing steps have an obvious impact on fidelity as well.

3 Big Data

As anticipated in Sect. 1, the term "big data" refers to structured or unstructured data sets that are impossible to store and process using common software tools (e.g., relational databases), regardless of the computing power or the physical storage at hand. Typically, *volume*, *velocity*, and *variety* are used to characterize the key properties of big data. They are the so-called three V's of big data:

- Volume refers to the size of the data;
- Velocity refers to the data provisioning rate and to the time within which it is necessary to act on them. Every minute about 400.000 tweets on Twitter are posted, 200 millions of e-mails are sent, and 2 millions of Google search queries are submitted [40];
- Variety refers to the heterogeneity of data acquisition, data representation, and semantic interpretation.

To extract value and make big data effective, the importance of a fourth V of big data, i.e., *veracity*, is increasingly being recognized. Veracity directly refers to inconsistencies and data quality problems: With the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of data is far from perfect.

According to a *classification* proposed by *UNECE (United Nations Economic Commission for Europe)* [49], there are three main types of data sources that can be viewed as big data: *human sourced* (e.g., blog comments), *process mediated* (e.g., banking records), and *machine generated* (e.g., sensor measurements), cf. Table 4 for a summary. In the following, we adopt such a classification in order to discuss the meaningfulness of big data quality, showing complexity of defining unique concepts for all possible types of sources. Before, we describe each type of data source in detail.

Table 4 Main characteristics of UNECE data sources

Source	Structure	Human influence
Human sourced	Loosely structured	Direct
Process mediated	Structured	Indirect (e.g., data entry activities)
Machine generated	Well structured	None

In a spectrum, human-sourced data are the less structured data and machine-generated data are the more structured data. Process-based data have mixed characteristics of human-sourced and machine-generated data

3.1 Human-Sourced Information Sources

This information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio, and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are often ungoverned. This includes a vast amount of data types such as: social networks (Facebook, Twitter, LinkedIn, etc.), blogs and comments, Internet searches on search engines (Google, etc.), videos loaded in the Internet (YouTube, etc.), user-generated maps, picture archives (Instagram, Flickr, Picasa, Google Photos, etc.), data and contents from mobile phones (text messages, etc.), e-mails, and so on.

3.2 Process-Mediated Sources

Business processes record and monitor business events of interest, such as registering a customer, manufacturing a product, and taking an order. The process-mediated data thus collected include transactions, reference tables, and relationships, as well as the metadata setting the context. Traditional business data are the vast majority of the information managed and processed by information technologies, in both operational and Business Intelligence (BI) systems. Process-mediated data are usually structured and stored in relational database systems. Examples include: data produced by public bodies and institutions (medical records, etc.), and data produced by the private sector (commercial transactions, banking/stock records, e-commerce, credit cards, etc.).

The so-called *deep Web* [3,25][3] is perhaps the most notable process-mediated sources of big data. It includes the contents hidden behind HTML forms, such as banking/stock records, e-commerce, and medical records. In order to get to such content, a user has to submit a form filled in with valid input values and is therefore difficult for search engines to index it. It represents a large fraction of the structured data on the Web, and it has been a long-standing challenge for the database community [28,44,54].

3.3 Machine-Generated Sources

Machine-generated sources leverage the impressive growth in the number of sensors and machines used to measure and record the events and situations in the physical world (cf. the IoT—Internet-of-Things [48] and CPS—Cyber-Physical Systems [53], trends). The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed are beyond traditional approaches. Examples include: data from fixed sensors (building automation sensors, weather/pollution sensors, traffic sensors/web cameras, scientific sensors, security/surveillance videos/images, etc.), data from mobile sensors, i.e., for tracking or analysis purposes (satellite images, GPS, mobile phone locations, car devices, etc.), and data from computer systems (log files, Web logs, etc.).

4 Big Data Quality

Given the variety of big data, a quality characterization of them should be *source specific*. Source specificity is most evident when considering the heterogeneous nature of some sources. For instance, data streams from a sensor network can be quality-characterized by the fact that data are often missing, and when not missing, they are subject to potentially significant noise and calibration effects. In addition, because sensing relies on some form of physical coupling, the potential for faulty data is high. Depending on where a fault occurs in the data reporting, observations might be subject to unacceptable noise levels (e.g., due to poor coupling or analog-to-digital conversion) or transmission errors (packet corruption or loss). Conversely, for social media data, data are highly unstructured, and often not accompanied by metadata. This means that high percentages of these data cannot be simply used by automated processes as they are affected by high percentages of noise. In the other cases, however, dedicated and often expensive activities of semantic extraction must be performed.

This is basically the thesis of this work: Big data quality in the broad term is a meaningless concept, as it should be defined in source-specific terms and according to the specific dimension(s) under investigation (as discussed in Sect. 2). In addition, the definition of such dimensions for big data, even if inspired by the traditional ones (discussed in Sect. 2), is quite complex due to unstructuredeness of data, semantics, etc., and still target of active research nowadays. In the following, in order to give the reader the intuition of such a complexity, we describe, for an example case of a data source type in Table 4 defined by UNECE, a comprehensive set of source-specific dimensions. We start with the most structured one and we analyze then the less structured sources (i.e., human sourced).

4.1 Process-Mediated Sources

A process-mediated data source provides a subset of objects in a particular domain and values of a subset of attributes for each object. An object usually possesses multiple types of data. For example, for health data, a patient's record includes age, height, weight, address, and measurements. Process-mediated data are distributed on the Web, and for each domain, there are many sources that needs to be integrated and fused together, for providing a high- quality representation of the real-world underlying process.

Data quality of Web structured data is discussed in [10, 12,36,37,45]. In the following, we describe a data model for process-mediated data sources and we suggest a set of source-

Table 5 Clusters of dimensions for process-mediated big data sources

Cluster	References
Accuracy, Reliability	[12,36,37]
Consistency	[37]
Redundancy	[37]
Spread, Value of the tail, Connectivity	[10,37]
Copying	[37]
Freshness, Coverage	[45]

Dimensions in *italic* are representative of the cluster

[3] Unfortunately, media and press created confusion about deep Web and *dark Web*, being the latter a (very) small portion of the deep Web that has been intentionally hidden and is inaccessible through standard Web browsers. The most famous content that resides on the dark Web is found in the TOR network, i.e., an anonymous network that can only be accessed with a special Web browser, called the TOR browser. This is the portion of the Internet most widely known for illicit activities because of the anonymity associated with the TOR network. In the following, we use the deep Web in a proper way, not to include the dark part of it. Cf. http://www.brightplanet.com/2014/03/clearing-confusion-deep-web-vs-dark-web/, http://brightplanet.com/wp-content/uploads/2012/03/12550176481-deepwebwhitepaper1.pdf (both accessed August 2015).

specific dimensions, including the metrics used in [10,12,36,37,45]. See Table 5 for a summary.

Data Model Given a set of sources in a particular domain (e.g., flights), we consider objects of the same type, each corresponding to a real-world entity (e.g., an object in the flight domain can be a particular flight on a particular day). Entities can change dynamically over time, i.e., new entities may appear and disappear, or the values of existing entities may change. For each object, we consider a set of attributes (e.g., scheduled departure time and actual departure time). For each attribute, which we call *data item*, we assume that a single true value exists that reflects the real world (e.g., the actual departure time of a flight is the minute that the airplane leaves the gate on the specific day).

In order to assess the quality of a specific set of sources, we consider all the values provided for each data item. If the provided values are exactly the true values, the quality is high. Conversely, if the provided value is very different from the true values, the quality is low. Causes of low quality include:

- outdated values
- incomplete values
- conflicting values
- wrong values
- noise in the data extraction.

Notation Let S be a source. Let d be a data item and v the value provided by a given source, $V(d)$ be the set of different values provided on d by all the sources, v^* be the true value of d, $S(d)$ be the set of sources that provide values on d, and $S(d, v)$ be the set of sources that provide value v on d. Let $A(S)$ be the set of global attributes that S provides. Entities in a source at a time point t are classified in three sets:

- up-to-date, $Up(S, t)$, including the entities that also exist in the real world and have their attribute values in agreement with the world;
- out-of-date, $Out(S, t)$, including the entities for which the latest value changes are not captured by the source;
- nondeleted, including all the remaining entities, i.e., entities that have disappeared from the real world.

4.1.1 Redundancy

If there are many different provided values on the same data item, the set of sources is redundant. Metrics of redundancy include:

- *redundancy on objects* is the percentage of sources that provide a particular object;
- *redundancy on data items* is the percentage of sources that provide a particular data item.

4.1.2 Consistency

If many sources provide the same values for the same data items, the set of sources is consistent. Metrics of consistency include:

- *number of values* is the number of different values provided on d, which is the size of $V(d)$.
- *entropy* is

$$E(d) = -\sum_{v \in V(d)} \frac{|S(d, v)|}{|S(d)|} \log \frac{|S(d, v)|}{|S(d)|} \tag{1}$$

  (the higher the inconsistency, the higher the entropy).
- *deviation* is

$$D(d) = \sqrt{\frac{1}{|V(d)|} \sum_{v \in V(d)} \left(\frac{v - v_0}{v_0}\right)^2} \tag{2}$$

  where v_0 is the value provided by the largest number of sources (it applies to data items with numerical values).

4.1.3 Accuracy

One commonly used approach to eliminate conflicts from inconsistent sources is to conduct majority voting, so that information with the highest number of occurrences is regarded as the correct answer. Due to copying, the value provided by most sources may not be the correct value. Source accuracy, which deals with the closeness of values to a golden standard, provides a valuable tool for weighting votes and improves the overall data quality. The *accuracy* cluster for process-mediated sources includes the *accuracy* and *reliability* dimensions.
Accuracy If the values provided for the same data item are correct and consistent over time, the data sources are accurate. Metrics of accuracy include:

- *source accuracy* is the fraction of values provided by the given source that are correct;
- *accuracy deviation*: let us denote by T the set of time points in a period, by $A(t)$ the accuracy of a source at a time $t \in T$, and by A' the mean accuracy over T, the accuracy deviation is

$$Dev(S) = \sqrt{\frac{1}{|T|} \sum_{t \in T} (A(t) - A')^2} \tag{3}$$

- *average accuracy* is the average source accuracy.

Reliability If the values provided by a data source are close to the gold standard, the source is reliable. Metrics of reliability include:

- *loss function* is defined based on the data type.
  - Categorical data: the most commonly used loss function is 0–1 loss in which an error is incurred if the value is different from the gold standard:

$$L(d) = \begin{cases} 1 & \text{if } v = v^* \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

  - Continuous data: The loss function should characterize the distance from the value to the gold standard with respect to the variance of values across sources. One common loss function is the normalized squared loss, which is defined as:

$$L(d) = \frac{(v^* - v)^2}{\text{std}(V(d))} \tag{5}$$

4.1.4 Copying

Copying is not to be confused with consistency and can be measured with respect to common elements among sources, such as object and attribute sets. Metrics of copying include:

- *schema commonality* is the average Jaccard similarity between the sets of provided attributes on each pair of sources

$$C = \text{avg}_{S,S'} \frac{|A(S) \cap A(S')|}{|A(S) \cup A(S')|} \tag{6}$$

- *object commonality* is the average Jaccard similarity but between the sets of provided objects;
- *value commonality* is the average percentage of common values over all shared data items between each source pair.

4.1.5 Spread

Process-mediated data are distributed on the Web, and collecting and crawling different sources providing data on the domain of interest is a necessary step toward a high-quality representation of the real world. The spread of a set of sources represents the complexity of such a step. The *spread* cluster for process-mediated sources includes the *spread* of the different sources, the *value of tail* sources, and *connectivity* dimensions.

- *spread*: If one only needs to identify and wrap a few top sites in order to build a comprehensive set of sources, the spread is low. A comprehensive set should also include some redundancy to overcome errors introduced by a single source;
- *value of tail*: If one needs to construct a comprehensive database, including the extraction of unpopular entities (i.e., relevant to a smaller group of users), the tail has high value;
- *connectivity*: If the data sources can be easily discovered by a bootstrapping-based Web-scale extraction algorithms (i.e., where one starts with seed entities, use them to reach all sites covering these entities, and iterate), the sources are connected.

4.1.6 Freshness

The freshness of a source represents its ability of reflecting real- world changes. The *freshness* cluster for process-mediated sources includes the *freshness* and *coverage* dimensions:

- the *freshness* of a source at a time t is the probability that a randomly selected entity is up-to-date, i.e.,

$$F(S) = \frac{|Up(S,t)|}{|S_t|} \tag{7}$$

  where S_t is the set of entities in the source at a time t;
- the *coverage* of a source is the probability that a random entity of the real world at a time t belongs to S, i.e.,

$$Cov(S) = \frac{|Up(S,t) \cup Out(S,t)|}{|W_t|} \tag{8}$$

  where W_t is the set of entities in the real world at a time t.

4.2 Machine-Generated Sources

A machine-generated data source measures and records the events and situations in the physical world. As sensors proliferate and data volumes grow, machine-generated data are becoming an increasingly important component of the information stored and processed by many businesses. Their well-structured nature is suitable for computer processing, but their size and speed are beyond traditional approaches.

Data quality of sensor data is discussed in [38,47]. In the following, we describe a data model for machine-generated data sources and suggest a set of source-specific dimensions, including the metrics used in [38,47] (see Table 6 for a summary).

Data Model A source in a particular domain (e.g., weather) provides discrete samples of real-world phenomena (e.g., wind). For each sample, we consider a set of attributes (e.g., speed and direction). For each attribute, which we call *data item*, we assume that a single true value exists that reflects

Table 6 Clusters of dimensions for machine-generated big data sources

Cluster	References
Accuracy	[38]
Completeness, Significance	[38]
Consistency	[38,47]
Trustworthiness	[38]
Freshness	[38]

Dimensions in *italic* are representative of the cluster

the real world (e.g., the actual speed is the speed of the wind on the specific day and time).

In order to assess the quality of a specific source, we consider the environment where the measures and records are taken (e.g., source location, measurement time, and source state) and the underlying measurement process. If the quality of such an environment and process is high, then the quality of the source is presumably high. Conversely, failures or malfunctions detected in the environment and process may lead to bad data. Causes of low quality include:

- hardware noise
- inaccuracies and impressions in sampling methods and derived data
- environmental effects
- adverse weather conditions
- faulty equipment.

Notation Let S be a source. Let d be a data item and v the value provided by a given source. $Lifetime(d)$ is the period of time after which a data item becomes obsolete and it is necessary to take a new value again. For example, the location of a fast moving vehicle may have a lifetime value smaller than the location of a walking person.

4.2.1 Accuracy

Source accuracy deals with the closeness of values to a golden standard and is directly affected by the measurement unit and the data type used. The location of an entity measured with the precision of ten meters is less accurate as compared to a measurement up to the precision of one meter. Metrics of accuracy include *precision*, i.e., the resolution of measurement unit of the sensor.

4.2.2 Completeness

This quality measure indicates the quantity of information that is provided by a source. The *completeness* cluster for machine-generated sources includes the *completeness* and *significance* dimensions.

Completeness If the set of attributes provided for a sample is exhaustive, then the completeness is high. Metrics of completeness include:

- *attribute ratio* is the ratio of the number of attributes available to the total number of attributes of the sample;
- *weighted attributes ratio* is the same as the attribute ratio, where the contribution of each attribute is proportional to its importance for the application of interest.

Significance The significance indicates the worthiness or the preciousness of a data item in a specific situation. Metrics of significance include the *critical value ratio*, defined as the fraction of an importance score of the data item for the application of interest, and the maximum importance score computed across all the data items.

4.2.3 Consistency

Sha and Shi [47] defines several subtypes of consistency, shown in Table 7, together with their definitions and an identification of whether the dimension refers to individual data or data streams. At a macrolevel, three types of consistency are considered, namely numerical, temporal, and frequency consistency. Notably numerical consistency is equivalent to accuracy; temporal consistency is meant as a degree of up-to-dateness; frequency consistency focuses on abnormal changes in data provisioning.

4.2.4 Trustworthiness

Trustworthiness of a source is highly affected by the distance between the sensor and the entity. The more a sensor is far away from the real-world entity, the more the correctness of information provided can be in doubt. *Trustworthiness* of a data item is defined as

$$T(d) = \begin{cases} (1 - \frac{dist}{d_{\max}} * \delta) & \text{if } d(s,e) < d_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $dist$ is the distance between the sensor and the entity, and $d_{\max}$ is the maximum distance for which we can trust on the observation of this sensor.

4.2.5 Freshness

The more a data item is fresh, the higher its validity of being used for a specific application at a given time. Metrics of freshness include:

- *age* of a data item, calculated by taking the difference between the current time, t_{curr}, and the measurement time of that data item $t(d)$;

- *up-to-dateness*: the up-to-dateness decreases as age increases, specifically

$$U(d) = \begin{cases} 1 - \frac{Age(d)}{Lifetime(d)} & \text{if } Age(d) < Lifetime(d) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

4.3 Human-Sourced Information Sources

Among the many human-sourced data sets, we specifically consider here the interesting case of short texts (e.g., Tweets, user queries in a search engine). We anticipate that the state of the art about quality of short texts is not very advanced as there are still very few approaches addressing it. Therefore, this section attempts to follow the structure of the previous ones, but without strictly adhering to it. Most of the considerations reported here are based on [27].

The major challenges in short text understanding are that short texts usually do not have the correct syntax that traditional POS-taggers or parsing methods can utilize and that they lack sufficient content to support statistical approaches to detect hidden topics. Furthermore, the vast amount of entity ambiguity also increases the difficulty of inferring the exact concepts. Humans can understand sparse, noisy, and ambiguous input such as short texts because they have knowledge of the language and the world. Many knowledge bases have emerged in recent years, including DBpedia, freebase, and Yago. Most of them are encyclopedic knowledge bases, containing facts such as Barack Obama's birthday and birthplace. They are essential for answering questions, but not for understanding them. To understand a question, knowledge of the language, e.g., the knowledge that birthplace and birthday are properties of a person, are needed; and lexical knowledge bases are constructed for this purpose. Hua et al. [27] uses a probabilistic lexical knowledge base known as Probase.

Data Model and Notation A short text is a text written in natural language with at most a dozen words. This includes

Table 7 Various types of consistency as defined in [47]

Types of consistency	Numerical/temporal/frequency	Individual data/data streams/both	Definition
Numerical	Numerical	Individual data	Collected data should be accurate
Temporal	Temporal	Individual data	Data should be delivered to the sink before or by it is expected
Frequency	Frequency	Both	Controls the frequency of dramatic data changes and abnormal readings of data streams
Absolute numerical	Numerical	Both	Sensor reading is out of the normal range, which can be preset by the application
Relative numerical	Numerical	Both	Error between the real field reading and the corresponding data at the sink
Hop	Numerical	Individual data	Data should keep consistency at each hop
Single path	Numerical and temporal	Individual data	Consistency holds when data are transmitted from the source to the sink using a single path
Multiple path	Numerical and temporal	Individual data	Consistency holds when data are transmitted from the source to the sink using multiple paths
Strict	Numerical and temporal	Data streams	Differs from hope consistency because it is defined on a set of data and requires no data loss
Alpha-loss	Numerical and temporal	Data streams	Similar to strict consistency except that alpha-data loss is accepted at the sink
Partial	Numerical and temporal	Data streams	Similar to alpha consistency except that temporal consistency is released
Trend	Numerical and temporal	Data streams	Similar to partial consistency except that numerical consistency is released
Range frequency	Frequency	Data streams	The number of abnormal readings exceed a certain number preset by the application
Change frequency	Frequency	Data streams	Changes of sensor readings exceed preset threshold

queries and microblogs. A term is a meaningful component of short text *s* which exists in a knowledge base, e.g., Probase. A role is a possible type of term; [27] considers two categories of roles, namely lexical roles (verb, adjective) and semantic roles (entity, concept, attribute). A typed term refers to a term *t* with role *r*. A concept vector expresses the semantics of an entity, where each element of the vector has the form <*c*, *w*>, where *c* is a concept in the knowledge base and *w* is the weight of the corresponding concept which can be obtained directly from the statistical information contained in the knowledge base. Finally, a concept cluster vector expresses the compressed semantics of an entity, where each element is a pair <*C*, *W*>, in which *C* represents a cluster of similar concepts and *W* is the weight sum of the contained concepts.

On the basis of the above model, an interesting dimension that can be defined for short texts is ambiguity.

4.3.1 Ambiguity

Hua et al. [27] distinguishes among three levels of ambiguities:

- Level 0 refers to entities that most people regard as unambiguous. These entities contain only one meaning, such as *dog* (animal), *California* (state), and *potato* (vegetable).
- Level 1 refers to entities that both make sense when treated as ambiguous or unambiguous. These entities usually have more meanings, but all of these meanings are related to some extent. For example, *Google* (company & search engine), *French* (language & country), *truck* (vehicle & public transportation) all belong to Level 1.
- Level 2 refers to entities that most people think as ambiguous. These entities have two of more meanings which are extremely different from each other, such as *apple* (fruit & company), *jaguar* (animal & company), *python* (animal & programming language).

Ambiguity of an entity can be computed through a statistical approach based on the previously cited vectors and knowledge bases (the reader can refer to [27] for details), and therefore, we can measure the ambiguity of a short text as the average ambiguity of the entities contained in it. Ambiguity of both entities and texts is in the range [0..1].

4.3.2 Other Dimensions

Clearly, many other dimensions, described in the previous sections, both specific for big data and for more traditional ones, can be applied to short texts. In particular, all the notions, metrics, and techniques described in Sects. 2.1.5, 2.1.7, and 2.1.8 (if we consider that Tweets can contain images) can be extended to this case.

5 Conclusion

In this paper, we have informally discussed the concept of data quality applied to big data, by highlighting how much complex is to define it in a general way. Data quality is already a multidimensional concept, difficult to characterize in precise definitions even in the case of well-structured data. Big data add two further dimensions to such complexity: (i) being "*very*" *source specific*, and for this we have adopted the UNECE classification, and (ii) being highly unstructured and schema-less, often without golden standards to refer to or very difficult to access.

In order to provide the reader the intuition of such complexities, after providing a tutorial section on data quality in traditional contexts (cf. Sect. 2), we have analyzed big data by providing insights into the UNECE classification (cf. Sect. 3), and then (cf. Sect. 4), for each type of data source, we have chosen a specific instance of such a type (notably deep Web data, sensors-generated data, and Twitters) and discussed how quality dimensions can be defined in such cases. The discussion shows how different data quality dimensions are for the three different cases.

Further work in the area is needed, especially for the case of human-generated data, in order to gain more insights into the concept of big data quality and its dimensions.

References

1. Batini C, Scannapieco M (2015) Data and information quality. Dimensions, principles and techniques. Springer, New York
2. Batini C, Palmonari M, Viscusi G (2012) The many faces of information and their impact on information quality. In: Proceedings of the 17th international conference on information quality (IQ 2012)
3. Bergman MK (2001) The deep web: surfacing hidden value. J Electron Publ 7:1407
4. Bizer C (2007) Quality-driven information filtering in the context of Web-based information systems, PhD thesis. Freie Universität Berlin, March 2007
5. Bizer C, Heath T, Berners-Lee T (2009) Linked data—the story so far. Int J Semant Web Inf Syst 5(3):1–22
6. Carroll JJ (2003) Signing rdf graphs. Technical report, HPL-2003-142, HP Labs
7. Chall JS (1995) Readability revisited. The new Dale-Chall readability formula, vol 118. Brookline Books, Cambridge
8. Cohen W, Ravikumar P, Fienberg S (2003) A comparison of string metrics for matching names and records. KDD Workshop Data Clean Object Consol 3:73–78
9. Crosby PB (1979) Quality is free. McGraw-Hill, New York
10. Dalvi N, Machanavajjhala A, Pang B (2012) An analysis of structured data on the web. Proc VLDB Endow 5(7):680–691

11. de Ridder H, Endrikhovski S (2002) Image quality is fun: reflections on fidelity, usefulness and naturalness. SID Symp Dig Tech Pap 33:986–989
12. Dong XL, Saha B, Srivastava D (2013) Less is more: selecting sources wisely for integration. In: Proceedings of the 39th international conference on very large data bases, PVLDB'13. VLDB Endowment, pp 37–48
13. DuBay WH (2004) The principles of readability. http://www.impact-information.com/impactinfo/readability02.pdf
14. Elmagarmid AK, Ipeirotis PG, Verykios VS (2007) Duplicate record detection: a survey. IEEE Trans Knowl Data Eng 19(1):1–16
15. Fan W, Geerts F (2012) Foundations of data quality management. Synthesis lectures on data management. Morgan & Claypool, San Rafael
16. Farr JN, Jenkins JJ, Paterson DG (1951) Simplification of flesch reading ease formula. J Appl Psychol 35(5):333
17. Fellegi IP, Holt D (1976) A systematic approach to automatic edit and imputation. J Am Stat Assoc 71(353):17–35
18. Flemming A (2011) Qualitätsmerkmale von Linked Data-veröffentlichenden Datenquellen. Diplomarbeit (Quality Criteria for Linked Data Sources) https://cs.uwaterloo.ca/~ohartig/files/DiplomarbeitAnnikaFlemming.pdf
19. Flesch R (1948) A new readability yardstick. J Appl Psychol 32(3):221
20. Fürber C, Hepp M (2011) Swiqa—a semantic web information quality assessment framework. In: Proceedings of the ECIS
21. Gal A (2015) Big data integration. In: Keynote speech at international conference on open and big data (OBD 2015), August 2015, IEEE CS Press
22. Gil Y, Artz D (2007) Towards content trust of web resources. Web Semant 5(4):227–239
23. Gonzales RC, Woods RE (2008) Digital image processing. Prentice Hall, Englewood Cliffs
24. Gunning R (1952) The technique of clear writing. McGraw Hill International Book, New York
25. He B, Patel M, Zhang Z, Chang K (2007) Accessing the deep web. Commun ACM 50(5):94–101
26. Hogan A, Umbrich J, Harth A, Cyganiak R, Polleres A, Decker S (2012) An empirical survey of linked data conformance. J Web Semant 14:14–44
27. Hua W, Wang Z, Wang H, Zheng K, Zhou X (2015) Short text understanding through lexical-semantic analysis. In: Poster at ICDE 2015
28. Ipeirotis PG, Gravano L (2002) Distributed search over the hidden web: hierarchical database sampling and selection. In: Proceedings of the 28th international conference on very large data bases. VLDB Endowment, pp 394–405
29. International Organization for Standardization - ISO. Quality management and quality assurance. Vocabulary. ISO 84021994
30. Jacobi I, Kagal L, Khandelwal A (2011) Rule-based trust assessment on the semantic web. In: International conference on Rule-based reasoning, programming, and applications series, pp 227–241
31. Juran JM (1988) Juran on planning for quality. The Free Press, New York
32. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document
33. Kitson HD (1921) The mind of the buyer: a psychology of selling, vol 21549. Macmillan, New York
34. Klare GR (1974) Assessing readability. Read Res Q 10:62–102
35. Lei Y, Uren V, Motta E (2007) A framework for evaluating semantic metadata. In: Proceedings of the 4th international conference on knowledge capture, ACM
36. Li Q, Li Y, Gao J, Zhao B, Fan W, Han J (2014) Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data
37. Li X, Dong XL, Lyons K, Meng W, Srivastava D (2012) Truth finding on the deep web: is the problem solved? Proc VLDB Endow 6(2):97–108
38. Manzoor A, Truong HL, Dustdar S (2008) On the evaluation of quality of context. In: Smart sensing and context. Springer
39. Mendes P, Mühleisen H, Bizer C (2012) Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 joint EDBT/ICDT workshops
40. NASSCOM (2012) Big data—the next big thing. Technical report, NASSCOM (2012)
41. Payne RS, McVay S (1971) Songs of humpback whales. Science 173:585–597
42. Pernici B, Scannapieco M (2003) Data quality in web information systems. J Data Semant 1:48–68
43. Pipino LL, Lee YW, Wang RY (2002) Data quality assessment. Commun ACM 45(4):211–218
44. Raghavan S, Garcia-Molina H (2001) Crawling the hidden web. In: Proceedings of the 27th international conference on very large data bases
45. Rekatsinas T, Dong XL, Srivastava D (2014) Characterizing and selecting fresh data sources. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data
46. Senter RJ, Smith EA (1967) Automated readability index. Technical report, DTIC Document
47. Sha K, Shi W (2008) Consistency-driven data quality management of networked sensor systems. J Parallel Distrib Comput 68(9):1207–1221
48. Stankovic JA (2014) Research directions for the internet of things. IEEE Internet Things J 1:3–9
49. UNECE. Classification of types of big data. http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data. Accessed Aug 2015
50. W3C. http://www.w3.org/WAI/. Accessed Aug 2015
51. Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12(4):5–34
52. Wayne SR (1983) Quality control circle and company wide quality control. Qual Prog 16(10):14–17
53. Wu FJ, Kao YF, Tseng YC (2011) From wireless sensor networks towards cyber physical systems. Pervasive Mobile Comput 7(4):397–413
54. Wu W, Yu C, Doan A, Meng W (2004) An interactive clustering-based approach to integrating source query interfaces on the deep web. In: Proceedings of the 2004 ACM SIGMOD international conference on management of data
55. Zakaluk BL, Samuels SJ (eds) (1988) Readability: its past, present, and future. International Reading Association, Newark