FROM DATA QUALITY TO BIG DATA QUALITY

Carlo Batini (*), Anisa Rula (*), Monica Scannapieco(-), Gianluigi Viscusi (+)
(*) University of Milano-Bicocca, Department of Informatics, Systems and Communication (DISCo)
(-)Istat - Italian National Institute of Statistics, Rome, Italy
(+) CDM (MTEI-CSI), École Polytechnique Fédérale de Lausanne

(*) batini@disco.unimib.it, rula@disco.unimib.it
(-)scannapi@istat.it
(+) gianluigi.viscusi@epfl.ch

Abstract

This article investigates the evolution of data quality issues from traditional structured data managed in relational databases to Big Data. In particular, The paper examines the nature of the relationship between Data Quality and several research coordinates that are relevant in Big Data, such as the variety of data types, data sources and application domains, focusing on maps, semi-structured texts, linked open data, sensor & sensor networks and official statistics. Consequently a set of structural characteristics is identified and a systematization of the a posteriori correlation between them and quality dimensions is provided. Finally, Big Data quality issues are considered in a conceptual framework suitable to map the evolution of the quality paradigm according to three core coordinates that are significant in the context of the Big Data phenomenon: the data type considered, the source of data, and the application domain. Thus, the framework allows ascertaining the relevant changes in data quality emerging with the Big Data phenomenon, through an integrative and theoretical literature review.

Keywords: big data, data quality, big data quality

INTRODUCTION

The area of Big Data (BD) is currently subject of intense investigation in academic literature, pushed by the growth of data made available in the Web and collected by fixed and mobile sensors. According to (Dumbill, 2013) "*Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it*".
Another issue that in recent years raised the attention of scholars and practitioners is Data Quality (DQ), a multifaceted concept, to the definition of which different dimensions concur. Data quality has been investigated focusing especially on data as represented in the relational model, traditionally adopted in Data Base Management Systems (for an extensive survey of DQ in the relational model see Batini & Scannapieco, 2006), notwithstanding the growing relevance and concerns of non-standard data such as text, music, design information and pictures (Rose, 1991). More recently, a variety of data types rising from linguistic and visual information, used and diffused through social networks, enterprise and public sector information systems as well as the Web, resulted in a deep investigation on how data quality concepts can be extended to such vast set

of data types, encompassing, e.g., semi-structured texts, maps, images, linked open data. Thus, the information growth consequent to the BD phenomenon has deeply impacted on the diversity of available types of data, the proliferation of sources of data, and the consequent great expansion of application domains.
Taking the above issues into account, in this paper we investigate how the multifaceted issues making up DQ have evolved from the traditional domain of databases to the domain of BD. The first coordinate we chose to analyze the evolution of the DQ concept are data types adopted in BD. In particular, we will analyze semi-structured texts, maps, and linked open data. Then, we will consider two other coordinates: (ii) the sources that originate BD, and (iii) application domains in which Big Data are used/investigated. As to sources, we will focus on sensors & sensor networks and as to application domains, we will focus on official statistics.
The article is organized as follows. First, we describe the methodology followed in the paper, that adopts an integrative review perspective for a theoretical purpose. Then we present the conceptual framework for analyzing the evolution of the DQ issues from relational databases to the diverse data types, application domains and sources considered in the following. As for DQ issues, we consider dimensions classified in terms of dimensions clusters, adopting the clusters proposed in Batini, Palmonari, and Viscusi (2012). The three BD coordinates, namely data types, sources and application domains are analyzed in terms of their structural characteristics. Subsequently, the evolution paths dealt with in the paper are introduced. Every path considers the evolution of a dimensions cluster from the relational domain to the issues target of the BD coordinates above introduced (i.e., data types, sources and application domains), further showing how the evolution of a given dimension can be interpreted a posteriori according to the structural characteristics considered. A final general discussion on DQ dimension clusters and BD coordinates concludes the paper.

METHODOLOGY

The article adopts an integrative review perspective (Beyea & Nicoll, 2015; Torraco, 2005; Whittemore & Knafl, 2005), aiming to summarize what is actually known on DQ that can provide insights on how to face the challenges of BD quality. In particular, the focus is on the evolution of data quality dimensions. The need for this review is motivated by the emergent nature of BD quality, that is more than the sum of its parts, exemplified by the data types, data sources and application domains analyzed in the subsequent sections. Consequently, these parts make up the conceptual framework guiding the analysis of the evolution of quality in BD, together with the key constructs resulting from a classification activity (Bailey, 1994) on the corpus of papers considered in the literature review. Thus, besides considering the insights by Webster & Watson (2002), we discuss the different steps followed in our literature review, adopting the streams of activities discussed by Boell & Cecez-Kecmanovic (2014). In particular, we focus here on what they call the inner hermeneutic cycle, made up of searching, sorting, selecting, acquiring, reading, identifying, refining. It is worth noting that an initial corpus of about 1.600 papers as well as related tables and notes has been included as basis for the literature review. The corpus resulted from a former literature review on DQ carried out by two of the authors of this paper from May 2013 to December 2014. Consequently, the subsequent searching activity has been informed by the knowledge of the two authors and information coming from their literature review.
Starting on February 1st, 2015 and involving all the authors of this paper, the searching activity has been carried out on databases for different research areas (information systems, information science, computer science, among others), such as Scopus, Web of Science, IEEE Explore, ACM Digital Library, Informs, AIS Electronic Library (AISeL). The keywords and search operators used are as follows: data quality OR information quality, data quality AND Big Data, information quality AND Big Data, quality AND Big Data. As for the sorting activity, we have first considered the presence

of the keywords in the title or in the abstract, the number of citations (when available in the search engine otherwise using Google scholar as a proxy), having a minimum thread of 100 citations, then we have analyzed papers for the period 2005 – 2014. However, it is worth noting that before moving to the time-related sorting we checked the abstract and the body of text for samples of 10 papers having less than 100 citations. As for the selection process, besides an analysis of title, abstract, and keywords, an additional activity of citation tracking has been conducted in order to check the completeness of the corpus of papers actually considered. During the acquiring step, when documents were not available we evaluated their relevance and content, reading the citations referring them in other papers before proceeding at buying or borrowing it from other institutions. Finally, as for the reading and identifying activity, while having a common shared Dropbox folder, the four authors of this paper have adopted different methods and tools for their own analysis, spanning from keeping notes on a text document to or using tools as NVIVO. Thus, this activity has required a weekly skype call to align the different perspectives and contributions. The different understandings emerging from the calls have guided the refining activity through additional searches using, e.g., domains- (e.g., official statistics) and source-related keywords (Internet of Things, IoT, crowdsourcing, sensors, etc.), leading to the consolidation of the corpus considered for this article actually being made up of 600 papers. As for this number, a final remark is worth pointing out that it includes also former relevant reviews and books on data even if published before 2005 (such as, e.g., Redman, 1996; Strong, Lee & Wang, 1997; Wand & Wang, 1996; Wang, Storey, & Firth, 1995). Finally, for theoretical purpose, in this paper we use only a summary of the literature review results to provide a conceptual framework for analyzing the evolution of quality in BD.

CONCEPTUAL FRAMEWORK FOR ANALYZING THE EVOLUTION OF QUALITY IN BIG DATA

Data quality dimensions and dimension clusters in detail

A common classification framework characterized by several quality dimensions, allows us to compare dimensions across different data types. The framework is based on a classification in clusters of dimensions proposed by Batini et al. (2012) where dimensions are included in the same cluster according to their similarity. Clusters are defined in the following list, where the first item in italics is the representative dimension of the cluster, thus introducing other member dimensions:

1. *Accuracy*, correctness, validity and precision focus on the adherence of data 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. *Redundancy*, minimality, compactness and conciseness refer to the capability of representing the reality of interest with the minimal use of informative resources.
4. *Readability*, comprehensibility, clarity and simplicity refer to ease of understanding of data by users.
5. *Accessibility* and availability are related to the ability of the user to access data from his or her culture, physical status/functions, and technologies available.
6. *Consistency*, cohesion and coherence refer to the capability of data 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.
7. *Trust*, including believability, reliability and reputation, catching how much data derive from an authoritative source.

Data types

We consider three main types of data that can be viewed as BD: (i) *maps*, (ii) *semi-structured texts* and (iii) *linked open data.* Each data types has inherently associated a set of structural characteristics, which are relevant for the investigation of quality dimensions defined in the literature, as shown in Table 1. In this section we discuss in detail the specific structural characteristics for the three data types.

Table 1. Structural characteristics for maps, semi-structured text, linked open data.

Data type	***Structural characteristics***
Maps	- *Space topological* - *Space geometric* - *Space thematic* - *Temporal*
Semi structured text	- *Lexical* - *Syntactic* - *Semantic* - *Rhetorical* - *Pragmatic*
Linked Open Data	- *Dereferenceable Resource* - *SPARQL Endpoint* - *RDF dump* - *Interlinking* - *Licensing*

Maps

A map can be defined as a representation, usually on a flat surface, of the features of an area of the earth or of a portion of the sky, showing them in their respective forms, sizes, and relationships (according to some convention of representation) and in their evolution in time. Maps are used for a vast amount of activities such as sailing or driving. Properties of data used to represent maps can be classified according to their structural reference to space, time, as well as the thematics (or themes) of the real world in their space localization and in their time evolution. Considering space, we can distinguish at least two types of properties of spatial objects: (i) *topology* and (ii) *geometry*. According to Schick (2007), topology is defined as "*the study of qualitative properties of certain objects (called topological spaces) that are invariant under a certain kind of transformation (called a continuous map), especially those properties that are invariant under a certain kind of equivalence (called homeomorphism)*". Thus, topology is a major area of mathematics concerned with the most basic properties of space, such as connectedness, continuity and boundary. Whereas geometry is the branch of mathematics concerned with issues of shape, size, relative position of figures, and the properties of space. Consequently, we are going to adopt a classification for quality dimensions of maps according to the target structured characteristics shown in Table 1, being them space-topological, space-geometric, space-thematic, and temporal.
Considering now *space-topological* and *space-geometric* characteristics, the different concepts and related primitives involved in topological and geometrical characteristics can be represented by means of conceptual schemas, also called "application schemas" in the geographical information

system (GIS) literature (Encyclopedia of GIS, 2010; Fonseca, Davis, & Câmara, 2003). As to geometry, in the ISO standard 19107 geometric characteristics are of three types: primitive, aggregate, and complex. Geometric primitives provide all components needed to depict the shape and the location of user artifacts such as buildings, roads, intersections, bridges, networks of roads, railways networks, or else natural phenomena, such as rivers, lakes, seas, mountains.
The latter refer to *space-thematic* characteristics of a territory, for which a map provider can adopt further sets of symbols or text that result in wider sets of rules that can be enforced for the set; such characteristics can be represented in terms of further application schemas. Some of the applications schemas have been standardized in ISO 19107, such as the ones representing roads and bridges of a road network. Other domains have not been standardized so far; in this case the provider of the map may introduce, explicitly or implicitly, new objects and relationships by means of new user defined application schemas.
Finally, *temporal characteristics* represent a major issue for data quality in general and particularly for maps. Indeed, as pointed out by Guptill (1995), one of the main concerns here is related to ascertain whether the temporal information adequately describes a geographic phenomenon, every geographic feature having a temporal aspect. Also, features as, e.g., the elevation of a geodetic control station or the ones described at a high spatial resolution, may have different times inertia, which require different intervals of inspection or validation (Guptill, 1995).

Semi-structured text

In the context of this paper, a semi-structured text refers to a digital text that neither conforms to the formal structure of data models associated with relational databases nor is structured for computation by a machine through tags or metadata. Therefore, semi-structured texts include both the digitization and digitalization (Tilson, Lyytinen, and Sørensen, 2010) of natural language texts, spanning from conversations, to newspaper articles, comments, books, etc. These texts have however structural characteristics referring to the issues discussed in what follows.
When we use natural language, the sentences we write or pronounce are characterized by a *lexicon* (a catalogue of a language's words), and a *grammar,* establishing a set of structural rules for word composition in meaningful sentences. Grammar is made up of *morphology* (accounting for the internal structure of words), and *syntax* (describing how words are combined to form sentences). Thus, the first two characteristics we consider for semi-structured texts are *lexical* and *syntactic* characteristics (the latter being chosen as representative for grammar, while not considering here the morphological ones). Furthermore, due to meaning and contextual issues, other subfields of a structure-focused study of language are worth considering relevant. In particular, *semantics*, that provides the meaning of sentences, texts and collection of texts, and *rhetoric,* that concerns the use of language for an effective speaking or writing, exploiting figures of speech and compositional techniques; while *pragmatics* is the way in which context contributes to meaning. Consequently, in what follows we consider also *semantics*, *rhetorical* and *pragmatic* structural characteristics.

Open and Linked Open Data

The Web has been in the last years an extraordinary vehicle of production, diffusion, and exchange of information. Data, as the lowest level of abstraction from which information is actually derived, can be provided on the Web as open data under the open data initiative (http://globalopendatainitiative.org/). Open data are mainly provided in different domains including economy, science, employment, environment and education (see, e.g., the European Union Open Data Portal at https://open-data.europa.eu/en/data/). Open data gain popularity with the rise of the Internet and World Wide Web especially, with the launch of open-data government initiatives.
The philosophy behind open data has been long established in public bodies, while the term “open data” itself is recent. In Bauer and Kaltenbock (2012) the authors adopt the following set of

properties for open data: data must be complete, primary, timely, accessible, machine processable, non discriminatory, with non proprietary format, license free, while the Open Data Handbook (2015) provides a definition based on the "openness" in relation to data and content "*Open data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike".* Then, open data become open linked data when, according to Tim Berners-Lee (2006):

- Information is available on the Web (any format) under an open license.
- Information is available as structured data (e.g. an Excel sheet instead of an image scan of a table).
- Non-proprietary formats are used (e.g. CSV instead of Microsoft Excel).
- URI identification is used so that people can point at individual data.
- Data is linked to other data to provide context.

Linked data enable publishers to link and publish structured data by generating semantic connections among data sets. Linked data exhibit structural characteristics referring to the issues discussed in what follows.

As in the Web of documents, a Uniform Resource Identifier (URI) is used in linked data to identify a document describing an entity (i.e. a real world object or an abstract concept). The use of URIs through specific protocols such as the application level protocol, e.g., the Hypertext Transport (or Transfer) Protocol (HTTP) enables interoperability between independent information systems. Each URI identifying an entity can be dereferenceable through the HTTP mechanism. This mechanism returns the description of the entity in a specified data format and language indicated by a user agent. This characteristic refers to the *resource* mechanism.

Considering the machine-readable data characteristic, a standard language, called Resource Description Framework (RDF), represents the description of entities. The RDF representation of documents enables different applications to process the standardized content. Further, RDF data are made accessible through *SPARQL endpoints* by executing SPARQL queries or *RDF dumps*. Linked data distributed across the Web apply a standard mechanism for specifying the connections between real world objects, named *interlinking*. The mechanism of interlinking is provided through RDF links that enable the process of discovering, accessing and integrating data in a straightforward way.

Linked open data further includes an explicit license (Heath and Bizer, 2011) *Licensing* is defined as the granting of permission for a consumer to re-use a data set under defined conditions (Zaveri, Rula, Maurino, Pietrobon, Lehmann, Sören, 2015). A license enables information consumers to use the data under clear legal terms. The existence of a machine-readable license as well as a human-readable license are important not only for the permissions a license grants but also as an indication of target requirements the user has to meet.

Taking the above issues into account, in what follows we consider *Dereferenceable Resource*, *SPARQL Endpoint*, *RDF dumps*, *Interlinking*, and *Licensing* as structural characteristics of linked open data.

Sources

Sensors & sensor networks Big Data, both for scientific purposes and for "Web of data" usage, are captured by a variety of devices; among them, sensors & sensor networks are becoming the most pervasive. Sensor networks can be defined as large-scale ad hoc networks of homogeneous or heterogeneous, compact, mobile or immobile sensor nodes that are randomly deployed in an area of interest (Gallegos, 2010). Sensor nodes collect different types of data, e.g., application specific environmental parameters, meteorological or Global Positioning System coordinates. These data can be in different forms, digital or analogue, spatial or temporal, alphanumeric or image, fixed or moving. Recent advances in miniaturization and low-cost, low-power design have led to active research in large-scale, highly distributed systems of small, wireless, low power, unattended sensors

and actuators. The vision of many researchers (Elson, 2003) is to create sensor-rich "smart environments" through large-scale deployment of microprocessors into the environment, each combined with radios capable of short-range wireless communication and sensors, that can detect local conditions such as temperature, sound, light, or the movement of chemicals or objects. Thus, sensor data are transferred, merged, transformed, and aggregated in sensor networks to extract complex knowledge. Ad-hoc deployable, wireless sensor networks can observe the environment in a fundamentally different way than previous classes of systems over a wide area, and densely in both time and space. Wallis, Borgman, Mayernik, Pepe, Ramanathan and Hansen (2007) analyses span over several evolutions of sensor networks. Most applications of wireless sensing systems in the environmental sciences are static deployments: sensors are placed in appropriate positions to report data continuously on local conditions. Sensors are monitored, both by humans and by computers, to determine changes in conditions. Autonomous networks can rely on machine actuation to capture scientifically relevant data, to alter data collection (e.g., capture data more frequently if excessive pollution is suspected), or to report emergencies that require intervention (e.g., faults in dams, water contamination).

Table 2. Structural characteristics for sensor & sensor networks

Source	***First level characteristics***	***Second level characteristics***
Sensors & sensor networks	*Space and time*	- *Single sensor* - *Whole sensor network or parts of it* - *Time*
	Shape of data	- *Individual data* - *Data streams*

As a consequence of the above discussion, structural characteristics of sensors and sensor networks can be referred to two coordinates (see Table 2). A first coordinate is related to space and time. As to space, we may be interested in the quality of data at single sensors or in the whole network or subparts of it. As a second coordinate, quality can be valuated both for individual data and for data streams.

Application domains

Official statistics

The main purpose of official statistics is well-defined by Principle 1 of the Fundamental Principles of Official Statistics (OS), as provided by the United Nations Statistics Division (1994): official statistics provide an indispensable element in the information systems of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information.
The quality of data resulting from OS production by National Statistical Institutes is therefore a primary issue. National Statistical Institutes started investigating the roles that BD can have in official statistics either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers (UNECE, 2013). Recently, the Scheveningen memorandum (DGNIS, 2013), which has the role of providing strategic guidelines to European

national offices, clearly stated that, given the opportunities that BD offer to OS, National Statistical Institutes are encouraged to undertake initiatives to examine the potential of Big Data sources in that regard.

Table 3. Structural characteristics for official statistics application domain.

Application domain	Structural characteristics
Official statistics	- *Coverage and sampling-related* - *Design-Related* - *Schema-related* - *Estimates-related* - *Integration-related*

A number of issues related to BD are specific to the OS domain, thus worth to be considered structural characteristics of it, that are summarized in Table 3 and discussed in what follows.
Populations covered by BD sources are not typically the target populations of OS and are often not explicitly defined. Moreover, given that the BD generating mechanisms are not under OS control, data deriving from BD sources can be selective, i.e. not representative of the target population. Dealing with these issues is not easy, especially because it is not always feasible to assess the relationships between the covered population and the target population on the one hand, and to estimate the bias to control, on the other hand. Thus, we consider a first set of structural characteristics as related to *coverage and sampling*.
Furthermore, considering data processing, this issue is concerned with three key aspects for dealing with Big Data in OS, namely: (i) data preparation, (ii) data filtering, (iii) data reconciliation. With respect to (i), big sources are typically event-based rather than unit-based, as it traditionally happens for OS survey data (or for administrative data). Hence a first preparation step is needed in order to deal with such new types of data. With respect to (ii) data filtering, BD are often affected by "noise" as to the analysis purpose that must be considered. On the one hand, this noise is related to the fact that the data generation process is not under a direct control of the statistician, which cannot apply a design activity to the data collection phase. On the other hand, the noise can be related to the particular nature of some sources, like unstructured information sources (e.g. Twitter data). Consequently, we can consider *design-related* structural characteristics.
With respect to (iii), even when some schema or metadata information is present in BD sources, such metadata need to be reconciled with metadata driving the statistical production, hence a reconciliation step is needed. As a further observation, due to the great variety of schema information that can derive from BD sources (e.g., Internet data), the reconciliation step can be very hard, due to the sparsity/incompleteness of BD sources schemas. So, we consider these issues concerning *schema-related* structural characteristics.
However, data analysis approaches traditionally used within Official Statistics may not be directly applied to Big Data analysis. Methodologies that proceed by exploratory analysis, like those based on data mining and machine learning, could be, instead, more appropriately applied. However, they are new for OS: though they are currently successfully applied in specific domains (e.g. customer profiling), their usage in the OS domain has still to be properly investigated. In this case, we talk of *estimates-related* structural characteristics.
Finally, a relevant issue concerns the usage of BD sources integrated with survey-based data or administrative data sources. However, several problems have been identified: (i) linking BD is hard because of privacy issues that prevent BD vendors to release data that are identifiable; (ii) integration task requires a precise and explicit structural metadata representation (schema information) that is often not available for BD; (iii) even when schema information is available, it will need to be reconciled with traditional sources schemas. Accordingly, here *integration-related* structural characteristics emerge as relevant.

ANALYSIS OF EVOLUTION LINES

In this section we analyze how the structural characteristics of the three coordinates considered in the previous sections namely, data types (DT), data sources (S) and application domains (AD) have influenced the evolution of quality dimensions in the different dimension clusters. The evolution of dimensions cluster and corresponding coordinates considered refer to (see Figure 1):

- **DT1** *Accuracy* for maps (a data type).
- **AD** *Completeness* of official statistics (an application domain).
- **DT2** *Readability* for semi-structured data (a data type).
- **DT3** *Accessibility* for linked open data (a data type).
- **S** *Consistency* for sensors & sensor networks (a source).
- **DT4** *Trust* for linked open data (a data type).

It is worth noting that the redundancy cluster is the only dimension cluster not considered in our analysis. For each coordinate (type of data, source, application) and corresponding associated dimension cluster we first shortly discuss quality dimensions in the relational model and then we discuss the evolution of dimensions determined by structural characteristics. We refer the interested reader to the extended discussion on dimensions in the relational model in Batini and Scannapieco (2006).

----------Figure 1 here----------

Evolution of quality dimensions in data types

Maps

Accuracy in the relational model

Accuracy may refer to data or else to the schema. Accuracy of data may refer to their actual value or else to the accuracy of the update. Accuracy of the actual value refers to *syntactic* accuracy, when the value is compared to a definition domain, e.g. Crlo is incorrect since does not match with any first name; accuracy of the actual value refers to *semantic* accuracy when the value does not match with the true value, e.g. Carlo is incorrect since it does not match with the true value Carla.

Table 4. Accuracy quality dimensions of maps classified by maps structural characteristics.

	Structural characteristics			
	Space – topological	***Space – geometrical***	***Space – thematic***	***Temporal***
Quality dimension	**Quality sub-dimension**			
Accuracy		-Positional accuracy -Relative positional accuracy -Absolute positional accuracy -Gridded data positional accuracy -Horizontal positional accuracy -Vertical accuracy	-Thematic accuracy -Quantitative attributes accuracy -Non-quantitative attributes accuracy -Correctness of classification	-Temporal validity -Accuracy of time measurement

The temporal accuracy of the update is the time span between the event of change in the real world and the corresponding update in the database. Schema accuracy can be defined with reference to the correct usage of the model constructs or else to the correct representation of requirements in the data schema.

Accuracy in maps

In the discipline of Geographical information systems, the quality of maps has been investigated for a long time and standardization bodies have produced several standards focused on maps quality. Thus, in what follows we refer to the ISO 19100 series of geographic information standards (with a specific focus on the 19107 Geographic information – Spatial schema standard) as well as the Spatial Data Transfer Standard (Shi, Fisher and Goodchild, 2003).

Table 5. Definitions of accuracy map dimensions and related sources.

Quality dimension	Quality sub-dimension	Source	Definition
Accuracy	Positional	ISO 19100	Accuracy of the position of features
	Relative positional	ISO 19100	Closeness of the relative position of features in a dataset to their respective positions accepted as or being true
	Absolute positional	ISO 19100	Closeness of reported coordinate values to values accepted as or being true
	Horizontal positional	SDTS	Accuracy of the horizontal position in the dataset
	Vertical positional	SDTS	Accuracy of the vertical position in the dataset
	Gridded data position	ISO 19100	Closeness of gridded data position values to values accepted as or being true
	Thematic	ISO 19100	Accuracy of quantitative attributes and the correctness of non quantitative attributes and of the classifications of features and their relationships
	Of quantitative attributes	ISO 19100	Accuracy of quantitative attributes
	Temporal validity	Batini and Scannapieco (2006)	Currency of a data update with respect to the event of change in the real world
	Of a time measurement	ISO 19100	Correctness of the temporal reference of an item
	Of non quantitative attributes	ISO 19100	Correctness of non quantitative attributes
Correctness	Of classification	ISO 19100	Comparison of the classes assigned to features or their attributes to a universe of discourse

In Table 4 we classify quality dimensions for maps according to structural characteristics. We clearly see the evolution of dimensions influenced by the map characteristics. Syntactic and semantic accuracy are now differentiated according to all relevant topological, geometric, thematic and temporal characteristics. In *Table **5*** we provide definitions for accuracy map dimensions and related sources, see (ISO 19100) and the Spatial Data Transfer Standard (STSD).

Semi-structured texts

Readability in the relational model

Readability pertains to the characteristic of the schema to make the user comprehend with low effort the reality represented in the schema. In the relational model, readability has a proxy in normalization (Elmasri and Navathe, 2001) that corresponds in a semantic interpretation to the separation of different entities and relationships between them in different tables, e.g. different tables for Student, Course and Exam instead of one single table. In the Entity Relationship (ER) model, used in conceptual design of relational databases, readability is also intended as

diagrammatic readability, namely the property that in the ER diagram certain aesthetic criteria such as avoiding crossings among lines are respected.

Readability in semi-structured texts

We can classify relevant dimensions for texts according to lexical, syntactic, semantic, rhetorical and pragmatic structural characteristics. In Table 6 we see the quality dimensions we will consider in this section related to semi-structured texts.

Table 6. Quality dimensions of semi-structured texts classified by structural characteristics.

	Structural characteristics				
	Lexical	***Syntactic***	***Semantic***	***Rhetorical***	***Pragmatic***
Quality dimension	**Quality sub-dimension**				
Readability	- Lexical readability	- Syntactic readability			
	-Text comprehension -Closer-to-text base comprehension -Closer to situation model level comprehension				

Readability is defined as reading easiness, especially as it results from a writing style. The majority of metrics proposed for readability are based on factors that represent two broad aspects of comprehension difficulty: (i) word lexical features and (ii) sentence or syntactic complexity.
As a consequence of the above perspective, readability is usually measured by using a mathematical formula that considers lexical or syntactic features of a given text, such as word length, and sentence length (see Table 6). Over 200 formulas have been reported for readability in the English language (DuBay, 2004) from 20s to 80s of the last century, among them Gunning Fox index, Automated readability index, Flesch reading ease, Flesch Kincaid grade level.
The shortcomings of traditional formulas become evident when one matches them against psycholinguistic models of the processes that the reader brings to bear on the text. Psycholinguists consider reading as a multicomponent skill operating at a number of different levels of processing: lexical, syntactic, semantic, and discoursal (Koda, 2005); the latter corresponds to rhetorical and pragmatic levels in Table 6. Thus, a psycholinguistic based assessment of text comprehensibility must include measures of text cohesion as well as meaning construction and encode comprehension as a multilevel process (Koda, 2005).
As to *text comprehension*, besides the general dimension, two more specific levels of comprehension are considered:

- *closer-to-text base comprehension* can be operationally defined as performance on comprehension questions that require minimal information integration (i.e., information explicitly stated within a sentence);
- *closer-to- situation model level comprehension* is defined by performance on comprehension questions that require more extensive information integration (i.e., bridging that involves integration of information across two or more sentences).

Linked Open Data

As to linked open data, we discuss the evolution of two quality dimensions, namely accessibility and trust.

Accessibility in relational data

Accessibility measures the ability of the user to access data from his or her own culture, physical status/functions, and available technologies. 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, with specific concern on accessibility for disabled persons. Guidelines referring to relational tables in Web sites are provided by the World Wide Web Consortium through the Web Accessibility Initiative (W3C-WAI2015). The W3C guidelines (W3C2000) identify the characteristics of the HTML representation of tables to be made accessible by means of assistive technologies, for example:

- for all data tables, identify row and column headers;
- for data tables that have two or more logical levels of row or column headers, use markup to associate data cells and header cells;
- for data tables elements, label elements with the "scope", "headers", and "axis" attributes, so that future browsers and assistive technologies will be able to select data from a table by filtering on categories.

Accessibility in linked open data

Public bodies, for reasons of transparency and accessibility, have progressively published public data in order to enable citizens to access data for their own purposes and interests. To make the data accessible in a standard way, the first step is to release the format of data from proprietary formats to open formats (i.e. RDF), which are not only understood by humans, but also by machines. The format issue is considered in several structural characteristics discussed for linked open data in the previous section, corresponding to several possible mechanisms that can be adopted to improve accessibility. In Table 7 we classify the relevant quality dimensions according to such mechanisms.

Table 7. Quality dimensions of linked open data classified by linked data structural characteristics.

	Structural characteristics				
	Dereferenceable Resource	***SPARQL Endpoint***	***RDF dumps***	***Interlinking***	***Licensing***
Quality dimension	***Quality sub-dimension***				
Accessibility	Resource accessibility	Dataset accessibility	Browsing accessibility	Integration accessibility	Reuse accessibility

One mechanism can be the use of HTTP URI, a combination of globally unique identification (through URIs) and a retrieval mechanism (through HTTP), which enables the identification of objects and abstract concepts and their descriptions; in this case the accessibility dimension refers to *dereferenceability*, or *resource accessibility*. To make datasets available through SPARQL endpoints, the user should indicate the URI of the dataset and the location of the corresponding SPARQL endpoint and should check whether the server responds to a SPARQL query; in this case we refer to *dataset accessibility*. A further mechanism to access a dataset is by making an RDF dump available for download; in this way the location of the RDF dump can be exploited, and we refer to *browsing accessibility*.

In order to specify the connection between real world objects, a mechanism of interlinking has been proposed based on the RDF links. Interlinking refers to the degree to which objects are linked to each other, be it within or between two or more data sources. It represents a relevant dimension for accessibility in linked data, since the process of data integration is made possible through the links created between various data sets. In this case the accessibility dimension corresponds to

integration accessibility, since RDF links describe the relationship between objects and enables discovering new data through integration.
Previous approaches to accessibility have evolved to investigate the new juridical licensing aspects of data. Licensing is a new quality dimension not considered for relational databases but mandatory in an open data world. Providing licensing information is an indication of how much data is accessible to be potentially re-used, based on the specification of legal rights and allowances; in this case the accessibility dimension corresponds to *reuse accessibility*.

Trust in databases
Relational databases are used in information systems of organizations as the basic technology for data management. In this context, trust is seen as the security characteristic that guarantees the authorized and reliable access to data by users and software applications. Areas interested by trust and security are:

- *access control*, that establishes which subjects (e.g. user, group) can perform which actions (e.g. read, write) on resources (e.g. a relational table, a column of a relational table);
- *integrity*, that aims at preventing unauthorized and improper data modification;
- *authentication*, the process by which the system verifies the identity claims of users.

Trust in linked open data
Contrary to what happens for relational databases, traditionally used inside the boundary of public or private organizations information systems, linked open data originates in the Web, through a variety of phenomena, often uncontrolled and undisciplined. This is also related to a general trend in data management to outsource data to 3rd party systems that would provide, for example, as a service functions such as data retrieval, with consequent privacy-preserving issues (see Kozak, Novak and Zezula 2014). Taking the above issues into account, trust in linked open data takes another sense, focusing on authority and reliability of the data provider. In this case, we do not identify a univocal correspondence between characteristics of linked open data and dimensions; rather we highlight several correlations, shown in Table 8. Correlations are defined between characteristics of (i) linked open data and (ii) the linked open data life cycle, represented in columns, and three dimensions proposed in the literature related to trust for linked open data, namely believability, verifiability, and reputation.

Table 8. Correspondence between structural characteristics of linked open data and new dimensions.

	Structural characteristics					
Quality dimension	***Provenance metadata***	***Metadata about the owner***	***Digital signature***	***Subjective opinions of consumers***	***Third party***	***Page ranks***
Believability	x	x		x	x	x
Verifiability	x		x		x	
Reputation		x		x	x	x

We discuss the three dimensions and related characteristics in more detail:

1) *Believability* refers to the subjective measure of a sure belief that data is true and credible. Believability can be measured as follows:
   a) assess the trustworthiness of RDF statements based on provenance information and on the opinion of data consumers;
   b) meta-information about the identity of information provider: checking whether the provider/contributor appears in a list of trusted providers;
   c) by a trusted third party which provides information such as citation count or page ranks.

2) *Verifiability* refers to the degree by which a data consumer can assess the correctness of the data set. It can be measured:
   a) by providing basic provenance information along with the dataset, such as using existing vocabularies like SIOC, Dublin Core, Provenance Vocabulary, the OPMV2 or the recently introduced PROV vocabulary:
   b) through the usage of digital signatures (Carroll, 2003), whereby a source can sign either a document containing an RDF serialization or an RDF graph:
   c) by an unbiased third party, if the dataset itself points to the source.
3) *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 dataset. There are different possibilities to evaluate 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 provider who publishes a data set. The (semi-) automated approach can be performed by the use of metadata on the owner, external links or page ranks. Finally, tracking of reputation is mentioned in the literature as also performed through a centralized authority.

Evolution of quality dimensions in data sources

Sensors & sensor networks

In this section we examine the evolution of the consistency cluster, with reference to sensor & sensor networks for the diverse structural characteristics of this specific kind of data source (see Table 9). In this case, dimensions are classified according to pairs of characteristics pertaining to (i) the space/time classification and (ii) the shape of data classification (see Table 2).

Table 9. Consistency sub-dimensions for sensors & sensors networks.

	Structural characteristics - 1			
Structural characteristics - 2	***Space – single sensor***	***Space – whole sensor network***	***Time***	***All of them***
Individual data	-absolute numerical error consistency	-relative numerical error consistency - hop consistency - single path cons. - multiple path cons.	- temporal consistency - frequency consistency	
Data streams	- α loss consist.		- partial - range frequency cons. - change frequency cons. - trend consistency	- strict consist.

Consistency in the relational model

The consistency dimension captures the violation of semantic rules defined over (a set of) data sets and related data items, where items can be tuples of relational tables or records in a file. With reference to relational theory, integrity constraints are an instantiation of such semantic rules. In statistics, data edits are another example of semantic rules that allow for the checking of consistency.

Consistency in sensors

Various types of quality dimensions for sensor & sensor networks (SN) are considered as subtypes of *consistency* (see, e.g., Sha 2010). The attention to consistency is due to the fact that sensors are almost never concentrated in a unique source of information, as, for instance, in a telescope, and are connected in networks with various topologies and graph schemes. Indeed, although a SN is an instance of a distributed system, there are several significant differences between them (Sha 2010). First, SNs are resource-constrained systems. Due to the memory size constraints and the large amount of sampled data, data is usually stored in sensors for a short period, and it will form data streams to be delivered to the sink(s) or base station(s). As a result, data consistency in SNs does not focus on the read/write consistency among multiple data replicas as in traditional distributed systems. Instead, data consistency in SNs is more interested in the *spatial* and *temporal consistency* of the same data, i.e. the consistency among several appearances of the data at different locations and at different times. So, space and time are intrinsic characteristics, as for maps, also for SNs. In this case, more than accuracy, space and time influence the consistency dimension cluster.
As a second point, SN applications usually operate on data streams, which can depict the trend of the parameters being monitored, or report a complex event. Thus, consistency models for data streams are more significant than those for individual data. Furthermore, compared with traditional distributed systems, the unreliable wireless communication is common, rather than exceptional, in SNs. Thus, in consistency models, the data loss resulting from unreliable wireless communication should also be considered.

Table 10. Types of consistencies relevant in sensor networks.

Types of Consistency	Definition
Absolute numerical error	The sensor reading is out of normal reading range, which can be pre-set by applications
Relative numerical error	The error between the real field reading and the corresponding data at the sink
Hop	The data should keep consistency at each hop
Single path	Holds when the data is transmitted from the source to the sink using a single path
Multiple path	Holds when the data is transmitted from the source to the sink using a multiple path
Strict	Differs from the hop consistency because it is defined on a set of data and requires no data lose
Temporal	The data should be delivered to the sink before or by it is expected
Frequency	Controls the frequency of data changes and abnormal reading of data streams in time
α -loss	Similar to strict consistency, except that α - loss data are accepted at the sink
Partial	Similar to α –loss consistency except that the temporal consistency is released
Range frequency	Detects if the number of outrange readings exceeds a pre-set maximum allowed number
Trend	Similar to partial consistency except that the numerical consistency is released
Change frequency	Detects if the number of dramatic changes in readings exceeds a pre-set threshold

The different types of consistency referring to SNs are shown in Table 10. Absolute numerical, relative numerical, hop, single path consistency refer to individual data correlated, in case of relative numerical and single path, at the reading sensor and at the sink, while in case of hop consistency is measured at a portion (the hop) of a signal's transmission from source to receiver. Multiple path, strict, temporal, frequency, α-loss, partial, and trend consistency refer to data streams. Multiple path consistency differs from single path since the whole network is considered; strict and α -loss consistency refer to completeness consistency, since they refer to absence of data loss, in the network or at the sink. Partial and trend consistency release previous dimensions in constraints referred to temporal and numerical characteristics of data.

Evolution of quality dimensions in application domains

Official statistics

Completeness in the relational model

The completeness of a relation characterizes the extent to which the table represents the corresponding real world's subject. Specific definitions for completeness can be provided by considering the granularity of the model elements, i.e., value, tuple, attribute and relation: (i) *value completeness* captures the presence of null values for some fields of a tuple, (ii) *tuple completeness* characterizes the completeness of a tuple with respect to the values of all its fields; (iii) *attribute completeness* measures the number of null values of a specific attribute in a relation, (iv) *relation completeness* captures the presence of null values in a whole relation.

Completeness in official statistics

Official statistics are an interesting Big Data domain, because of the emerging relevance of Internet data as complement, or actually as subject of experimentation (UNECE, 2013), for substituting traditional official statistics that are based on surveys questionnaires or administrative sources. Consequently, this complementarity to be effective requires a change in quality dimensions as known when applied to the relational model, also considered the methodological issues pointed out by UNECE (2013), such as, e.g., measures of quality of outputs produced from hard-to-manage external data supply. In what follows, the evolution of completeness is discussed with regard to sub-dimensions for the structural characteristics above examined and shown in Table 11. In particular, we are going to consider representativeness, selectivity, and sparsity.

Table 11. Quality dimensions considered for official statistics classified by structural characteristics.

	Structural characteristics				
	Coverage and sampling-related	***Design-Related***	***Schema-related***	***Estimates-related***	***Integration-related***
Quality Cluster	**Quality sub-dimension**				
Completeness	-Representativeness -Selectivity		-Sparsity		

Completeness is first challenged by sub-dimensions for coverage and sampling as well as design-related structural characteristics, that are *representativeness* and *selectivity*. As pointed out, e.g., by Buelens, Daas, Burger, Puts, and Van den Brakel (2014), a subset that is not representative is referred to as *selective*. Indeed, a subset of a finite population is representative of it with regard to a given variable, if the variable distribution within the subset is the same as in the population, otherwise is selective. As said above, given that the BD generating mechanisms are not under OS control, data deriving from BD sources can be selective, i.e. not representative of the target population. For example, as discussed by Buelens et al. (2014) social media data are selective because not all people in a given country post messages on social media platforms, and anyway at varying rates, and some accounts are managed by organizations and not by individuals.
Sparsity is another relevant sub-dimension, impacting schema-related, estimates-related, and, integration-related structural characteristics. Indeed, as above-mentioned, a great variety of sparse schema information can derive from Web data, with a consequent incompleteness of schema. Furthermore, this information has to be integrated and/or reconciled with what available in OS obtained through traditional controlled methods. In such a way, for example, it is possible to have aggregate figures on the sentiment in social media messages by people in a given country towards the current economic situation and OS statistics on consumer confidence (UNECE, 2013).

CONCLUSION

The paper has investigated the nature of the relationship between Data Quality and several research coordinates that are relevant in Big Data, such as the variety of data types, data sources and application domains, focusing on maps, semi-structured texts, linked open data, sensor & sensor networks and official statistics. We believe that the selected coordinates provide insights also for Big Data quality issues in areas such as business intelligence (see on the topic Chee, William, Shijia and Richards, 2014). In what follows we summarize the main results of the paper and possible areas of future research on this topic.
The variability and heterogeneity of coordinates typical of Big Data environments investigated in this paper has lead to:

- a classification of structural characteristics associated to each coordinate;
- a clustered classification of data quality dimensions;
- an a posteriori justification of the evolution of quality dimensions from relational data types in a database setting to dimensions mentioned in the literature for each coordinate.

The two topics of Data Quality and Big Data are both multifaceted, and, at the same time, are both characterized by a rapid evolution of paradigms considered at the state of the art. They are characterized by several similitudes in paradigms:

1. DQ can be investigated through the formalization of relevant dimensions (of quality), and related metrics; Big Data issues can be investigated in terms of structural characteristics, such as the variability of data types, sources of data and application domains;
2. the two areas need for the discovery of methods and techniques for the traditional life cycle of data: that for DQ corresponds to a) *collection*, b) *quality assessment*, and c) *improvement*; while for BD corresponds to a) *collection*, b) *fusion*, c) *analysis*, d) *processing*, and e) *usage*.

Although our investigation has covered only specific paths of the above mentioned evolution, we have achieved an in depth insight of a phenomenon that will inherently influence the future of BD quality. In order to be able to cope with the huge variability of methods and techniques needed to manage DQ in BD, we need to understand first the deep nature of the coordinates considered and then the correlation with dimensions adopted in methods and techniques. We notice that this is relevant for DQ as for the value and utility of BD, as well as other issues not considered here such as, e.g., filtering, integration and fusion of BD. Thus, a main result of the paper is a systematization of the *a posteriori* correlation between quality dimensions and structural characteristics. However, in order to fully achieve such objective, we have to extend the analysis to the whole dimensions (clusters) vs. structural characteristics matrix shown in Figure 1. A second area of future investigation refers to the a posteriori correlation between *metrics*, namely measurements associated to quality dimensions, and structural characteristics. Then, another long term more ambitious objective is the *a priori* discovery of relevant dimensions and metrics for a given BD coordinate. In this case, the target of the exploratory research launched in this paper is a methodological process that has as input (i) a quality dimension in a given quality dimension cluster, and (ii) a coordinate relevant in Big Data (data type, source of data, application domain) described in terms of its structural characteristics; such a methodological process should allow to discover or at least to explore the conception of specific dimensions and metrics, and possibly assessment and improvement methods and techniques for achieving Big Data quality.
Finally the evolution lines analyzed in this paper require a further investigation from a systemic perspective as the one adopted in Viscusi, Batini and Mecella (2010), arguing that each core cluster dimension may refer to other facets of information infrastructures than data, such as, e.g., legal frameworks, processes, services, communication networks, information systems.

REFERENCES

Bailey, K. D. (1994). *Typologies and taxonomies - An Introduction to Classification Techniques*. Thousand Oaks, CA: SAGE Publications.

Batini, C., Palmonari, M., & Viscusi, G. (2012) The many faces of information and their impact on information quality. In Équille, L.B., Comyn-Wattiau, I., Scannapieco, M. (Eds.): *Proc. 17th International Conference on Information Quality - ICIQ 2012* (pp. 212–228).

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Berlin Heidelberg: Springer-Verlag.

Bauer, F., & Kaltenböck, M. (2012) *Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers*. Vienna, Austria: edition mono/monochrom, ISBN: 978-3-902796-05-9

Berners-Lee, Tim. (2006, July 27) *Linked Data - Design Issues.* [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html.

Beyea, S. C., & Nicoll, L. H. (2015). Writing an integrative review. *AORN Journal*, *67*(4), 877–880.

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems*, *34*(1), 257–286.

Buelens, B., Daas, P., Burger, J., Puts, M., & Van den Brakel, J. (2014, March 28). *Selectivity of Big Data.* [Online]. Available: http://www.pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf.

Carroll, J. J. (2003). *Signing RDF Graphs*. In D. Fensel et al. (Eds.): *ISWC 2003*, LNCS 2870, pp. 369–384, 2003. © Springer-Verlag Berlin Heidelberg 2003.

Chee, C.-H., William Y., Shijia G., and Richards, G. (2014). Improving Business Intelligence Traceability and Accountability. *Journal of Database Management*, *25*(3), 28–47. doi:10.4018/jdm.2014070102.

DGNIS. (2013, September 27). *Scheveningen Memorandum - Big Data and Official Statistics*. [Online]. Available: http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13.

DuBay WH (2004, August 25). *The principles of readability.* [Online]. Available: www.impact-information.com/impactinfo/readability02.pdf.

Dumbill, E. (2013). Making Sense of Big Data (Editorial). *Big Data*, *1*(1), 1–2.

Elmasri, R., & Navathe, B. S. (2001). *Fundamentals of database systems*. Reading, Mass.: Addison-Wesley. ISBN: 0201542633.

Elson, J., & Römer, K. (2003). Wireless sensor networks: a new regime for time synchronization. *Computer Communication Review*, *33*(1),149-154.

Shekhar, S. & Xiong. (Eds.) (2008). *Encyclopedia of GIS*. Berlin / Heidelberg: Springer.

Fonseca, F., Davis, C., & Câmara, G. (2003). Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. *GeoInformatica*, *7*(4), 355–378.

Gallegos, I., Gates, A., Tweedie, C. (2010). DaProS: A data property specification tool to capture scientific sensor data properties. In J. Trujillo, G. Dobbie, H. Kangassalo, S. Hartmann, M. Kirchberg, M. Rossi, I. Reinhartz-Berger, E. Zimányi, F. Frasincar, (Eds.). *Advances in Conceptual Modeling – Applications and Challenges - ER 2010 Workshops ACM-L, CMLSA, CMS, DE@ER, FP-UML, SeCoGIS, WISM, Vancouver, BC, Canada, November 1-4, 2010. Proceedings - Lecture Notes in Computer Science Volume 6413* (pp. 232-241). Berlin / Heidelberg: Springer.

Guptill, S. C. (1995). Temporal information. In S. C. Guptill & J.L. Morrison (Eds.), *Elements of Spatial Data Quality* (pp. 153–165). International Cartographic Association. Oxford: Elsevier Science Ltd.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space* (1st ed.). San Rafael, Ca: Morgan & Claypool.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*.

Cambridge: Cambridge University Press

Kozak, S., Novak, D., Zezula, P.(2014). Privacy-Preserving Outsourced Similarity Search. *Journal of Database Management*, *25*(3), 48–71. doi:10.4018/jdm.2014070103.

Open Data Handbook (2015, October 1). *What is Open Data?*. [Online]. Available: http://opendatahandbook.org/en/what-is-open-data/

Redman, T. (1996). *Data Quality for the Information Age*. Norwood: Artech House.

Rose, E. (1991). Data Modeling for Non-Standard Data. *Journal of Database Management, 2*(3), 8–21. doi:10.4018/jdm.1991070102.

Sha, K. & Shi, W. (2008). Consistency-driven data quality management of networked sensor systems. *Journal of parallel and Distributed Computing*, *68*(9), 1207-1221.

Shi, W., Fisher, P., Goodchild, M. F. (Eds.) (2003). *Spatial data quality*. Boca Raton: CRC Press.

Shick, P. L. (2007). *Topology: Point set and geometric*. New York: Wiley & Sons.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data Quality in Context. *Communications of the ACM*, *40*(5), 103–110.

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Digital Infrastructures: The Missing IS Research Agenda. *Information Systems Research*, *21*(4), 748–759.

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review* , *4* (3), 356–367. doi:10.1177/1534484305278283.

UNECE. (2013, March 11). *What does Big data mean for official statistics?* [Online]. Available: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622.

United Nations Statistics Division. (1994, April 11-15). *Fundamental Principles of National Official Statistics*. [Online]. Available: http://unstats.un.org/unsd/dnss/gp/fp-english.pdf

Viscusi, G., Batini, C., & Mecella, M. (2010). *Information Systems for eGovernment: A Quality-of-Service Perspective*. Berlin / Heidelberg: Springer.

Wallis, J C., Borgman, C. Mayernik, M., Pepe, A., Ramanathan N. & Hansen, M. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. In L. Kovács, N. Fuhr, and C. Meghini (Eds.), *11th European Conference on Research and Advanced Technology for Digital Libraries September 16–21*, *Volume 4675 of the series Lecture Notes in Computer Science* (pp. 380-391). Berlin / Heidelberg: Springer.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95.

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A Framework for Analysis of Data Quality Research. *IEEE Transaction on Knowledge and Data Engineering*, *7*(4).

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MISQuarterly*, *26*(2), XIII–XXIII.

Whittemore, R., & Knafl, K. (2005). The integrative review: updated methodology. *Journal of Advanced Nursing*, *52*(5), 546–553. doi:10.1111/j.1365-2648.2005.03621.x

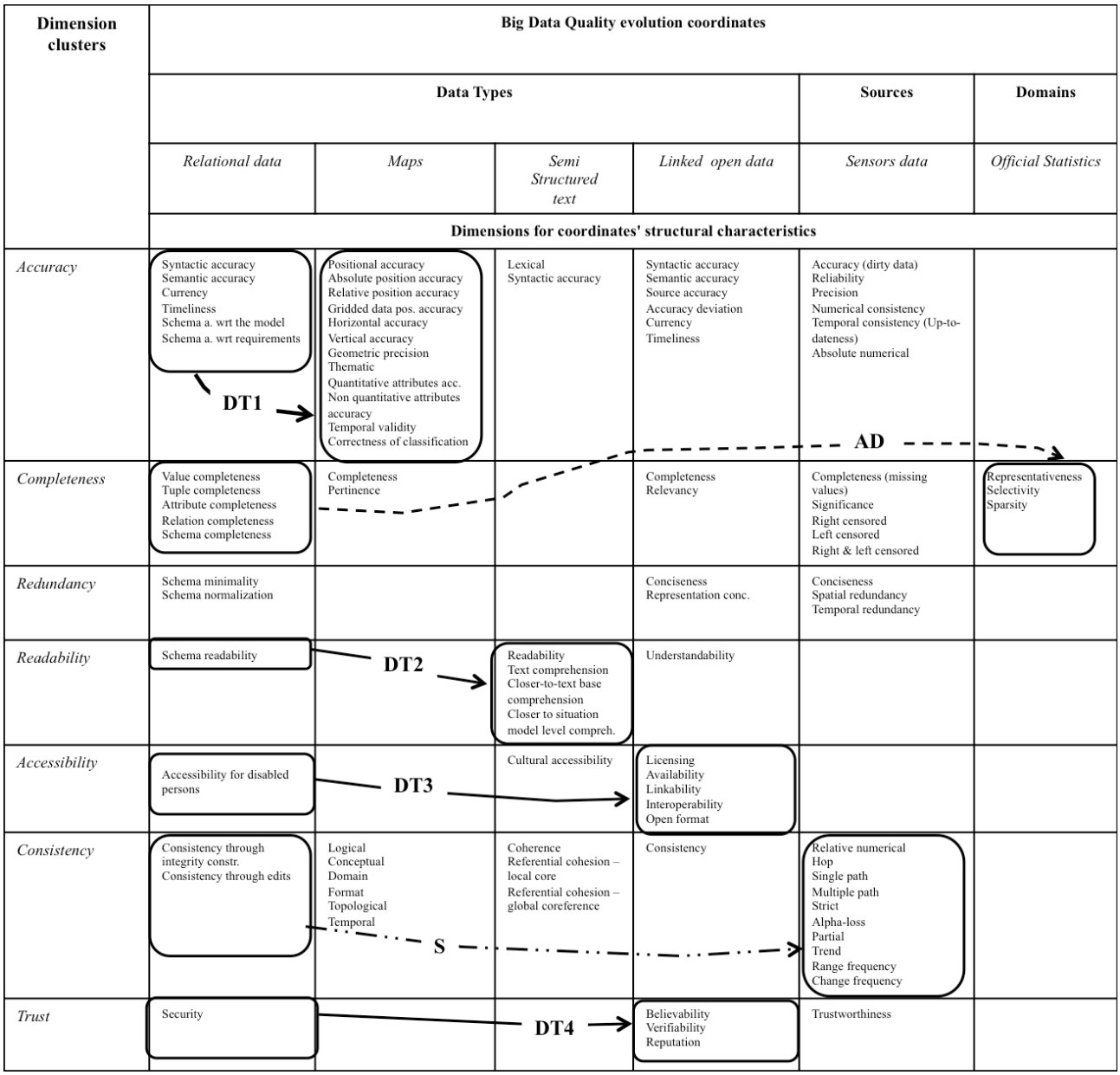
Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Sören, A. (forthcoming, 2016). Quality assessment for Linked Data: A survey. *Semantic Web Journal*, *7*(1), 1-32. [Online] Preview available: http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey

W3C-WAI (2015, March 19). *Web Accessibility Initiative (WAI) - Guidelines and Techniques.* [Online]. Available: http://www.w3.org/WAI/guid-tech.html.

W3C (2000, November 6), *HTML Techniques for Web Content Accessibility Guidelines 1.0*. [Online]. Available: http://www.w3.org/TR/WCAG10-HTML-TECHS/.

FIGURES

Figure 1. Dimension clusters and corresponding coordinates for which we analyze the role of structural characteristics.



Short biographies

Carlo Batini is full professor at University of Milan Bicocca. His main research areas have been in the past conceptual database design, schema integration, automatic layout of diagrams, visual query languages, repositories of conceptual schemas. More recently his interests have covered methodologies for e-Government and for the service life cycle, service portfolio management, information value and value of integration in databases and in service repositories. In 2013 he received in Hong Kong the Elsevier Peter P. Chen Award for his research, teaching and publishing activity in conceptual modeling.

Anisa Rula is a postdoc at the University of Milano-Bicocca. She received her PhD from the University of Milano-Bicocca in 2014. Her research interest is in providing methodologies and tools for measuring and improving Linked Data quality, with particular focus on time-related quality dimensions. She was program chair of 2nd LDQ Workshop at ISWC 2015 and she has been a reviewer of conferences and journals in the Semantic Web field including Semantic Web Journal and ISWC. In 2014 she received the Best Student Research Paper for the work entitled Hybrid Acquisition of Temporal Scopes presented at ESWC.

Monica Scannapieco is a researcher at Istat, the Italian National Institute of Statistics since 2006. She earned a University Degree in Computer Engineering with honors and a Ph.D. in Computer Engineering at SAPIENZA - Università di Roma. She is author of more than 100 papers mainly on data quality, privacy preservation and data integration published in the leading conferences and journals in databases and information systems areas. She has been involved in several European research projects on data quality and data integration.

Gianluigi Viscusi (PhD) is research fellow at the Chair of Corporate Strategy and Innovation (CSI) of the College of Management (CDM) at the École Polytechnique Fédérale de Lausanne (EPFL). His research interests include information systems planning, business modelling, public policy and technology innovation, e-Government, information quality and value, service management and engineering, social study of technology. Currently, his research focuses on crowd-driven innovation, social value of open government, and translational research in innovation and technology management. In 2010 he has co-authored with Carlo Batini and Massimo Mecella the book “Information Systems for eGovernment: a quality of service perspective” (Springer, Heidelberg).