Chapter 4
Opening the Closed World: A Survey of Information Quality Research in the Wild

Carlo Batini, Matteo Palmonari, and Gianluigi Viscusi

Abstract In this paper we identify and discuss key topics characterizing recent information quality research and their impact on future research perspectives in a context where information is increasingly diverse. The investigation considers basic issues related to information quality definitions, dimensions, and factors referring to information systems, information representation, influence of the observer and of the task. We conclude the paper by discussing how philosophical studies can contribute to a better understanding of some key foundational problems that emerged in our analysis.

4.1 Introduction

In the last decades, information systems of both private and public organizations have been migrating from a hierarchical/monolithic to a network-based structure, where the potential sources that single organizations or networks of cooperating organizations can use for the purpose of their activity is dramatically increased in size and scope. At the same time data representations have evolved from structured data, to semi-structured and unstructured text, to maps, images, videos and sounds. Now more than ever, information is available in different formats, media and resources and it is accessed and exploited through multiple channels. Each information is completely intertwined with the others, each contributing to the information assets of an organization. Among others, *data and information quality* (information quality in the following, IQ, for short), is becoming critical for human beings and organizations, referring to being able to define, model, measure and

C. Batini (✉) • M. Palmonari • G. Viscusi
Dipartimento di Informatica, Sistemistica e Comunicazione,
Università di Milano Bicocca, viale Sarca 336 – U14, 20037 Milan, Italy
e-mail: batini@disco.unimib.it; palmonari@disco.unimib.it; viscusi@disco.unimib.it

L. Floridi and P. Illari (eds.), *The Philosophy of Information Quality*,
Synthese Library 358, DOI 10.1007/978-3-319-07121-3_4,

improve the quality of data and information that are exchanged and used in everyday life, in business processes of firms, and administrative processes of public administrations.

However, it is our point that IQ issues are worth to be considered "in the wild", paraphrasing the title and the aims of the book by Hutchins (1995), where the terms "wild" referred to human cognition in its natural habitat, naturally occurring, and culturally constituted. As well as for cognition as investigated by Hutchins, we can consider the challenges and changes in the information quality paradigm when studied not only in the captivity of traditional database systems and IT units, but also in the everyday world of the information ecosystem produced by social networks and semantic information extraction processes. Accordingly, despite the relevance of the quality of information assets, the growing literature on information quality constructs and dimensions (Madnick et al. 2009; Wand and Wang 1996), we believe that a further clarification and formalization of their main concepts are required (Batini et al. 2012).

Thus, our aim in this paper is to make a comparative review of the recent literature on data and information quality, with the goal of providing several insights on recent developments along several dimensions. In Sect. 4.2 we critically discuss a recent standard that has been issued by the International Organization for Standardization (ISO). In Sect. 4.3 we introduce two coordinates that are used in the paper to survey the literature: *basic issues*, which concern founding features of IQ, and *influencing factors*, which represent aspects of information systems that have an influence on the interpretation and evaluation of information quality. Sections 4.4, 4.5 and 4.6 address the three basic issues, namely (a) IQ definitions (Sect. 4.4), (b) IQ dimensions (Sect 4.5), with specific reference to the accuracy and completeness dimensions, and c. IQ classifications (Sect. 4.6). Section 4.7 focuses on the relationships between IQ dimensions and the evolution of information systems, while Sects. 4.8 and 4.9 address the levels of semantic constraints and the evolution in the representation of data and knowledge from databases to web knowledge bases. Section 4.10 concludes the paper with a discussion focused on the relationships between IQ and philosophical issues.

4.2 Information Quality in the ISO Standardization Process

When attempting to formalize the concept of data and information quality, the first issue concerns the concepts of *data, information* and *quality*. Traditionally, international standard bodies are authoritative and knowledgeable institutions when definitional and classification issues are considered.

Luckily for our purposes, ISO has enacted in 2008 the standard ISO/IEC 25012:2008 (see Data Quality Model 2008), that defines data quality as the " degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions", and provides "a general data quality model for data retained in a structured format within a computer system". The document presents:

Table 4.1 Data quality characteristics in the ISO standard

DQ characteristic	Definition (all definitions except for completeness and accessibility begin with: the degree to which data has attributes that…")
Correctness	Correctly represent the true value of the intended attribute of a concept or event in a specific context of use
Completeness	Subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use
Consistency	Are free from contradiction and are coherent with other data in a specific context of use
Credibility	Are regarded as true and believable by users in specific context of use
Currentness	Are of the right age in a specific context of use
Accessibility	Data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability
Compliance	Adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use
Confidentiality	Ensure that it is only accessible and interpretable by authorized users in a specific context of use
Efficiency	Can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use
Precision	Are exact or that provide discrimination in a specific context of use
Traceability	Provide an audit trail of access to the data and of any changes made to the data in a specific context of use
Understandability	Enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use
Availability	Enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use
Portability	Enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use
Recoverability	Enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use

- a set of terms and definitions for concepts involved,
- two points of view that can be adopted when considering data quality *characteristics* (or *dimensions* (Batini and Scannapieco 2006), in the following),
  - the *inherent* point of view, that corresponds to intrinsic properties of data, and
  - the *system dependent* point of view, that depends on the system adopted to represent and mange data,
- a set of data quality characteristics and corresponding definitions, see Table 4.1.

When we look at the definitions of data and information proposed in the document, we discover that:

1. *data* is defined as "reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing",

2. *information* is defined as "information-processing knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts that within a certain context have a particular meaning".

This choice is specular to the usual one in textbooks and scientific papers, where information is defined in terms of data (see e.g. Floridi 2011), and knowledge in terms of information in some definitions (e.g. in Merriam Webster). The ISO effort shows severe limitations, such as:

1. the flat classification adopted among characteristics (see Table 4.1 for the list of characteristics proposed and corresponding definitions), that contradicts e.g. the classification provided in the document "ISO/IEC 9126 Software engineering — Product quality, an international standard for the evaluation of software quality", where quality characteristics are expressed in terms of sub-characteristics;
2. several characteristics (e.g. completeness) depend on the model adopted for data representation, even though this dependence is not explicitly discussed;
3. data organized in models that neatly distinguish between instances and schemas are considered, e.g. the relational model, while schemaless data, such as e.g. textual documents, are ignored;
4. there is no attempt to distinguish between different types of data and information, from structured data to texts and images.

As a consequence of the above discussion, we can consider the ISO standard as a first standardization effort of the concept of data quality, which needs further investigation and elaboration.

In the rest of the paper, when we refer to *data quality*, we make reference to quality of structured data, while when we refer to *information quality*, we consider wider types of data represented according to different heterogeneous models, such as semi-structured data, texts, drawings, maps, images, videos, sounds, etc. This pragmatic distinction reflects a common use of these terms in the technical literature.

4.3 Information Quality Research Coordinates: Basic Issues and Influencing Factors

We define two coordinates to better formalize and analyze several aspects of IQ. One coordinate is represented by IQ *basic issues* and another coordinate is represented by IQ *influencing factors*, which have been both defined in the introduction. In the following we list a set of items for each of these two coordinates; we do not claim that these items provide an exhaustive coverage of the two concepts; rather they have to be seen as a first attempt to characterize and classify the issues discussed in the literature on IQ, following a classificatorial approach similar to the one adopted in a previous analysis of data quality methodologies (Batini et al. 2009). The basic issues considered in this paper are:

BI1. Definitions of IQ – How many different definitions exist of information quality?

BI2. IQ Dimensions – How many dimensions are considered in the literature to capture the multifaceted character of the concept of IQ?

BI3. IQ dimension classifications – In how many ways dimensions can be classified?

A list of significant factors influencing IQ is:

IF1. Type of information representation – As investigated in Batini et al. (2008), types of information representation can change significantly: if we want to emphasize the *visual perceptual character* of information, we can consider images, maps, graphical representations of conceptual schemas; if we want to emphasize the *linguistic character* of information, we can consider structured, unstructured and semi-structured types of text (specific types of semi-structured text that have been considered in the literature are e.g. laws and medical records). Another common distinction is the one among *structured data*, i.e. data having a rigid and pre-defined schema like relational databases, *unstructured data*, i.e., data having no schema like images and texts in natural language, and *semi-structured data*, i.e., data with a schema that is unknown, flexible or implicit like data in XML. In addition to the above mentioned types of data, we also consider data represented with languages such as RDF and JASON (Antoniou and van Harmelen 2008), called *weakly structured* data in this paper, which have a basic structure (e.g., RDF data have a graph structure) but have non-rigid, possibly changing and third-party schemas attached to the data. Considering the diversity of data to be considered, does the type of information representation influence IQ?

IF2. Life cycle of information – Information has usually a life cycle, made of acquisition (or imaging), validation, processing, exchange, rendering and diffusion. Does the life cycle of the different types of information representations influence IQ?

IF3. Type of information system – Information system architectures have evolved from hierarchical systems, where the information is highly controlled, to distributed, cooperative, peer to peer, web based information, where information flows are anarchic and undisciplined. How this evolution has influenced IQ?

IF4. Level of semantic constraints: binding vs. freedom in coupling data and schemas and open vs. closed world assumption – Data can undergo different levels of semantic constraints. In databases, data and schemas are tightly coupled, while other data, e.g. RDF data, can be loosely coupled with schema level constraints by means of metadata. Moreover, the closed world assumption (CWA) usually holds in data bases, meaning that any statement that is not known to be true is false. In knowledge bases, the open world assumption (OWA) states that any statement that is not known, cannot be predicated neither true nor false. Do the binding/freedom in coupling schemas and data and CWA/OWA influence IQ?

IF5. Syntax vs. semantics – How the syntax vs. the semantics of information play a role in IQ?

IF6. Objective vs. subjective assessment of IQ – With the term subjective we mean "evaluated by human beings", while the term objective means "evaluated by a measurement performed on real world phenomena". How the *objective vs. subjective quality evaluation* is related with IQ?

IF7. Influence of the observer – How IQ is influenced by the observer/receiver, human being vs. machine?

IF8. Influence of the task – IQ is intrinsic to information or it is influenced by the application/task/context in which information is used?

IF9. Topological/geometrical/metric space in visually perceived information – How the different spaces influence IQ?

IF10. Level of abstraction of information represented – The same real world phenomenon can be represented at different levels of abstraction (see Batini et al. 1993) where levels of abstractions are defined for conceptual database schemas). To give a simple example, a conceptual schema in the Entity Relationship model made of the two entities `Student` and `Course` and the relationship `Exam`, can be abstracted in terms of a schema made of the unique entity `Exam`, having as identifier the couple of identifiers of `Student` and `Course` in the refined schema. Is IQ influenced by (e.g. changes of) the level of abstraction?

IQ is a relatively new discipline in information sciences. As a consequence, a discussion on above basic issues and influencing factors can be made at the state of the art in terms of examples and counterexamples leading to observations, statements, conjectures that cannot be formally stated and validated. Conscious of these limitations and immaturity, in the rest of the paper we discuss (some) basic issues, influencing factors and relevant relationships between them.

4.4 Definitions of IQ

We first deal with one of the most controversial questions around IQ: is there an intrinsic information quality? Look at Fig. 4.1, and before reading the next paragraph, reply to this question: which is the most accurate/faithful image of Mars? Perhaps you said: the first one on the left…

The first image has been downloaded from a blog, while the second from the NASA site. Your judgments were probably based on your own model of Mars. Now that you have some ancillary data you could change your opinion…So, may we come to the conclusion that an intrinsic information quality does not exist? This conclusion seems too strong if we look at the two images of Fig. 4.2; they seem to represent the same flower, it is hard to say that the image on the left is of good quality.

The two previous examples show that in order to predicate the quality of a piece of information, sometimes (Fig. 4.1) we need a reference version of the information, other times we evaluate the quality according to perceptual and/or technological

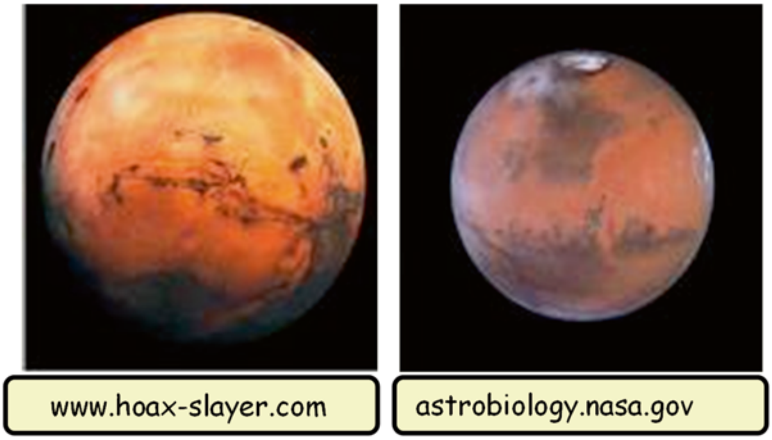


Fig. 4.1 Two pictures of Mars

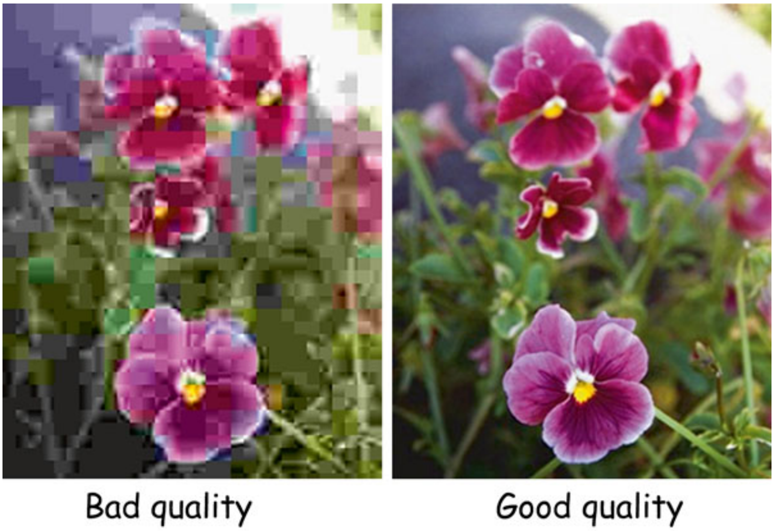


Fig. 4.2 Two images of flowers

characteristics of information, that depend on the type of information representation (IF1), such as, in this case, the image resolution, that can be measured subjectively or else in terms of a metrics based on dots per inch.

As another example, Fig. 4.3 shows five different version of a photo, that make use of a decreasing number of dots per inch; looking at the 7 Kb version, we consider acceptable the rendering of the image with respect to the original, while in the 2 K case the resolution is not perceived as acceptable. So, we can conceive a

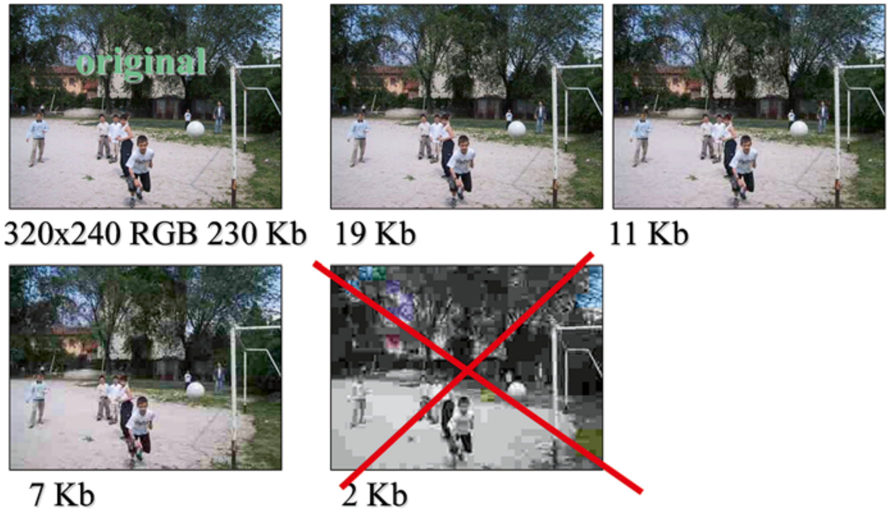


Fig. 4.3 Several representation of the same photo with decreasing amount of dots

concept of *minimal amount of data needed to represent a piece of information* over a threshold of reasonable quality. However we also observe that the context of use plays a role in defining this threshold; as an example, an image used as a web thumbnail is expected to be displayed at lower size (dpis and pixels) than the same image as a picture in a newspaper. We want now to investigate more in depth (see Table 4.2) the relationship between definitions of IQ in the literature and corresponding influencing factors shown in column 1 of the table.

Looking at columns, three different information representations are considered, (a) structured data, (b) images and (c) a specific type of semi-structured text, laws. We can define the quality of the image as the lack of distortions or artifacts that reduce the accessibility of its information contents. Some of the most frequent artifacts considered are: blurriness, graininess, blockiness, lack of contrast and lack of saturation. The definition referring to quality as a list of properties (BI2) is inspired by former contributions from the conceptual modeling research area (Lindland et al. 1994). Whereas the overall framework in Table 4.2 assumes the definition of data and information quality as based on the role of an information system as a representation (Wand and Wang 1996), and the consequent distinction between the internal and external views of an information system (Wand and Weber 1995). The internal view is use-independent, supporting dimensions of quality as intrinsic to the data; while the external view considered the user view of the real world system (the observer perspective), where possible data deficiencies happen (Wand and Wang 1996). Moreover, it is worth noting that most of the research effort in the literature on data quality has provided by far greatest attention to the design and production processes involved in generating the data as the main sources of quality deficiencies (Wand and Wang 1996). Notice also that the definition more closely influenced by

Table 4.2 Definitions of IQ and related issues and factors mentioned in definition

IF1 type of InfoR → Related issues/factors	Structured data	Images	Structured text: laws
IF2/IF6 Absence of defects		A perfect image should be free from all visible defects arising from digitalization and processing processes	
Adherence to the original			
BI2 – Quality as a list of properties	1. High quality data is accurate, timely, meaningful, and complete 2. The degree of excellence of data. Factors contributing to data quality include: the data is stored according to their data types, the data is consistent, the data is not redundant, the data follows business rules, the data corresponds to established domains, the data is timely, the data is well understood		
IF6/IF7 Impression of the observer		Impression of its merit or excellence as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted (III Association 2007)	

(continued)

Table 4.2 (continued)

IF1 type of InfoR → Related issues/factors	Structured data	Images	Structured text: laws
IF8 Fitness for use/adequacy to the task	Data are of high quality “if they are fit for their intended uses in operations, decision making and planning	1. The perceptually weighted combination of significant attributes (contrast, graininess, etc.) of an image when considered in its marketplace or application 2. Degree of adequacy to its function/goal within a specific application field	Laws whose structure and performance approach those of “the ideal law”: – It is simply stated and has a clear meaning – It is successful in achieving its objective – It interacts synergistically with other laws – It produces no harmful side effects – It imposes the least possible burdens on the people
Conformance…	…to requirements	Of match of the acquired/reproduced image with IF2 the original → fidelity IF7 viewer’s internal references → naturalness	

the observer (third row) claims for a “third party” subjective evaluation, not influenced by the domain.

Coming to the fourth row, we see that *fitness for use*, that corresponds to IF9, Influence of the task, is the only common driving issue, while the impression of the observer (IF6) is typical of images, that are characterized by a high prevalence of subjective measures on objective ones (IF7). According to IF9, IQ can be expressed quantifying how it influences the performance of the task that uses it. Focusing on images (Batini et al. 2008):

- In the framework of medical imaging, an image is of good quality if the resulting diagnosis is correct.
- In a biometric system, an image of a face is of good quality if the person can be reliably recognized.
- In an optical character recognition system a scanned document has a good quality is all the words can be correctly interpreted.

Finally we comment the conformance definition, which in case of images may be associated:

(a) to the original, focusing in such a way on possible distortions during the processing life cycle (IF2), as a consequence subsuming the possibility to access to the original (Ciocca et al. 2009; Gasparini et al. 2012), or else
(b) to viewer’s internal references (IF8), i.e. the perceived model in the user’s mind of the image (Ciocca et al. 2009; Gasparini et al. 2012).

This last characteristic is typical of information representations such as images, that may influence emotions of human beings (Ciocca et al. 2009; Gasparini et al. 2012).

4.5 IQ Dimensions

Many possible dimensions and metrics can be conceived for IQ. Focusing on structured data in data bases, 13 methodologies for the assessment and improvement of data quality are listed in Batini et al. (2009), which mention a total of about 220 different dimensions with repetitions and about 70 without repetitions. In Batini and Scannapieco (2006) several examples of synonyms and homonyms existing in the literature among dimensions are shown.

Focusing on most frequently mentioned dimensions, namely accuracy, completeness, consistency, timeliness, currency, in Table 4.3 we see that multiple metrics are defined for each dimension, some of them objective and others subjective (IF6).

Coming to specific dimensions, we now investigate more in depth accuracy and completeness.

Table 4.3 Dimensions and related metrics

Dimensions	Name	Metrics definition
Accuracy	Acc1	Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one Syntactic accuracy = number of correct values/number of total value
	Acc2	Number of delivered accurate tuples
	Acc3	User survey – questionnaire
Completeness	Compl1	Completeness = number of not null value/total number of values
	Compl2	Completeness = number of tuples delivered/expected number
	Compl3	Completeness of web data = (T_{max} – $T_{current}$)* (completeness$_{max}$ – completeness$_{current}$)/2
	Compl4	User survey – questionnaire
Consistency	Cons1	Consistency = number of consistent values/number of total values
	Cons2	Number of tuples violating constraints, number of coding differences
	Cons3	Number of pages with style guide deviation
	Cons4	User survey – questionnaire
Timeliness	Time1	Timeliness = (max (0;1-currency/volatility))
	Time2	Percentage of process executions able to be performed within the required time frame
	Time3	User survey – questionnaire
Currency	Curr1	Currency = time in which data are stored in the system – time in which data are updated in the real world
	Curr2	Time of last update
	Curr3	Currency = request time – last update
	Curr4	Currency = age + (delivery time – Input time)
	Curr5	User survey – questionnaire

4.5.1 Accuracy Dimension

Several methodologies investigated in Batini et al. (2009), see accuracy from two different points of view, syntactic and semantic (IF5). Figure 4.4 shows Italian first names, and compares them with the item "*Mrio*" that does not correspond to any of them. Semantic accuracy of a value *v* can be intuitively defined as closeness of the value *v* to the true value *v**; for a formal definition in the context of relational databases, the first order logic interpretation of the relational model can be adopted. Since semantic accuracy can be complex to measure and improve, a second type of accuracy, syntactic accuracy, measures the minimal distance between the value *v* and all possible values in the domain *D* of *v*. In our case, if we consider as distance the edit distance, the minimum number of character insertions, deletions, and replacements to convert "*Mrio*" to a string in the domain, the syntactic accuracy of "*Mario*", is 1. Notice that the string corresponding to "*Mrio*" is "*Mario*", but it could be possible that two errors have occurred so that the true value of "*Mrio*" is "*Maria*", another valid Italian name. To recognize this, we need more knowledge on the object represented by "*Mrio*", e.g. that is a male, or a female.

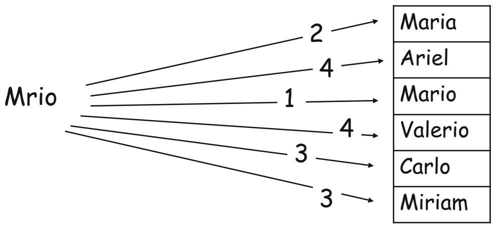


Fig. 4.4 Example of accuracy evaluation

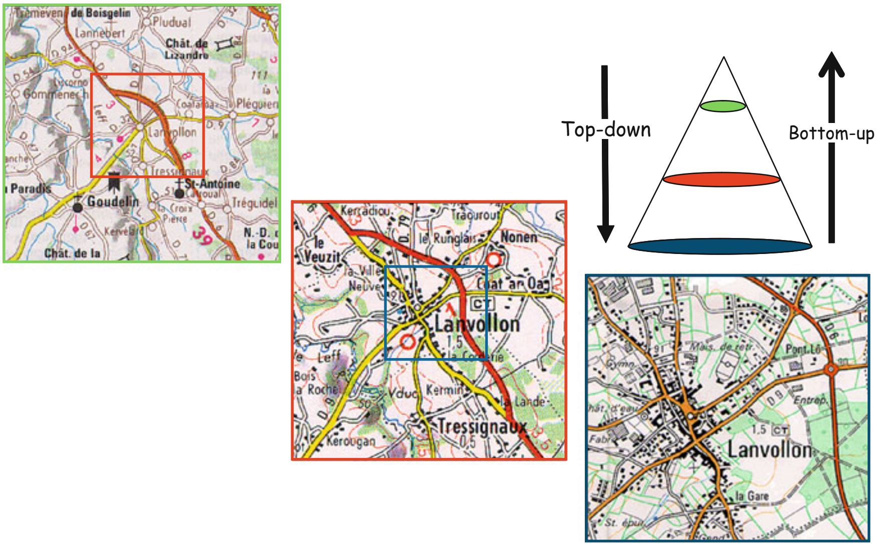
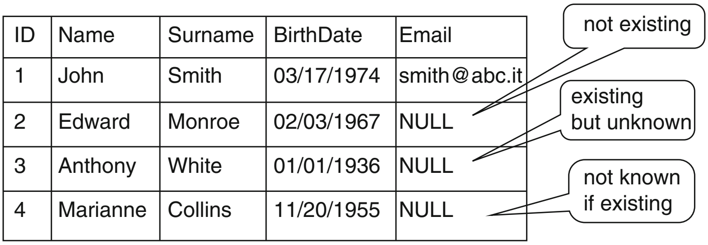


Fig. 4.5 The same geographic area represented at three abstraction levels

Another intriguing relationship to be investigated concerns accuracy and level of abstraction (IF10). Here we focus on maps. In our experience of visiting a city or making a travel by car, we need maps at different levels of detail. Cartographic generalization involves symbolizing data, and applying a set of techniques that convey the salient characteristics of that data. These techniques seek to give prominence to the essential qualities of the feature portrayed, e.g. that buildings retain their anthropogenic qualities – such as their angular form. In Fig. 4.5 we show the same geographic area around the town of Lanvollon in France represented at three abstraction levels.

As said in Encyclopedia of GIS (2010), "Different combinations, amounts of application, and different orderings of these techniques can produce different yet aesthetically acceptable solutions. The focus is not on making changes to information contained in the database, but to solely focus upon avoiding ambiguity in the interpretation of the image. The process is one of compromise reflecting the long held view among cartographers that making maps involves telling small lies in order to tell the truth!".



ID	Name	Surname	BirthDate	Email
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL

not existing
existing but unknown
not known if existing

Fig. 4.6 Completeness in relational tables

These considerations show that even a dimension such as accuracy, that is considered only from the inherent point of view in the ISO standard, is strongly influenced by the context in which information is perceived/consumed.

4.5.2 Completeness Dimension

The definition of completeness depends on the type of information representation (IF1), and is also influenced by the CWA/OWA (IF4). Let us consider the table reported in Fig. 4.6, with attributes `Name`, `Surname`, `BirthDate`, and `Email`. If the person represented by tuple 2 has no e-mail, tuple 2 is complete. If the person represented by tuple 3 has an e-mail, but its value is not known then tuple 3 presents incompleteness. Finally, if it is not known whether the person represented by tuple 4 has an e-mail or not, incompleteness may or may not occur, according to the two cases. In a model such as the relational model, in which only one type of null value is defined, these three types of incompleteness are collapsed into one.

Further, relation completeness, i.e., the number of tuples w.r.t. to the total number of individuals to be represented in the table, depends on the validity of the CWA or else of the OWA. Usually it is assumed that the closed world assumption holds in data bases, in this case a relation is always complete. Instead semantic data are usually considered under the OWA; if we adopt this assumption for our table, then we cannot compute completeness, unless we introduce the concept of reference relation, i.e. a relation that is considered complete, and used as a reference for measuring the completeness of other relations representing the universe, for details see Batini and Scannapieco (2006).

4.6 IQ Dimension Classifications

Several classifications of dimensions are considered in the literature, we shortly mention them, while their comparison is outside the scope of the paper. In Lee et al. (2002), a two ways classification is proposed based on

(a) conforms to specification vs. meets or exceeds consumer expectations (here we find an influence from IF6),
(b) product quality vs. service quality.

Wang and Strong (1996) proposes an empirical classification of data qualities, based on intrinsic, contextual, representations, accessibility qualities. The approach of Liu et al. (2002), is based on the concept of evolutional data quality, where the data life cycle is seen as composed of four phases:

- *Collection*, data are captured using sensors, devices, etc.
- *Organization*, data are organized in a model/representation.
- *Presentation*, data are presented by means of a view/style model.
- *Application*, data are used according to an algorithm, method, heuristic, model, etc.

Qualities that in other approaches are generically attached to data, here are associated to specific phases, e.g. accuracy to collection, consistency to organization. A theory in Liu et al. (2002) is a general designation for any technique, method, approach, or model that is employed during the data life cycle. E.g. when data in the Organization phase is stored, a model is chosen, such as a relational or object-oriented model to guide the data organization. Due to the attachment of data to theories, when defining quality, we need to consider how data meet the specifications or serve the purposes of a theory. Such a concept of quality is called *theory-specific*; e.g., in the relational model, theory specific qualities are normal forms and referential integrity.

In order to investigate the influence of the type of information representation (IF1) on the analysis of several quality dimensions, we use adopt in the following the classification of dimensions proposed in Batini et al. (2008), where dimensions are empirically included in the same cluster according to perceived similarity. Clusters concern:

1. *Accuracy/correctness/precision* refer to the adherence to a given reference reality.
2. *Completeness/pertinence* refer to the capability to express all (and only) the relevant aspects of the reality of interest.
3. *Currency/volatility/timeliness* refer to temporal properties.
4. *Minimality/redundancy/compactness* refer to the capability of expressing all the aspects of the reality of interest only once and with the minimal use of resources.
5. *Readability/comprehensibility/usability* refer to ease of understanding and fruition by users.
6. *Consistency/coherence* refer to the capability of the information to comply with all properties of the membership set (class, category,…) as well as to those of the sets of elements the reality of interest is in some relationship.
7. *Credibility/reputation*, information derives from an authoritative source.

In Table 4.4 we relate dimensions cited in the literature with the above dimension classification (BI3) and with a set of types of information representation (IF1).

Table 4.4 Comparative analysis of quality dimensions for diverse information representations

Quality dimension cluster	Structured data	Geographic maps	Images	Unstructured texts	Laws and legal frameworks
Correctness/accuracy/precision	**IF4** Schema accuracy w.r.t requirements w.r.t. the model **IF4** Instance accuracy **IF5** Syntactic **IF5** Semantic **IF8** Domain dependent (ex. Last Names, etc.)	Instance **IF9** Spatial accuracy Relative/absolute Relative inter layer Locally increased r.a. External/internal Neighbourhood a. Vertical/horizontal/ height Attribute accuracy **IF8** Domain dependent accuracy (ex. traffic at critical intersections, urban vs rural areas, etc.) Accuracy of raster representation	**IF8** Accuracy Syntactic Semantic "Reduced" semantic Genuineness Fidelity Naturalness Resolution Spatial resolution **IF2** Scan type	IF8 Accuracy IF5 Syntactic IF5 Semantic IF4 Structural similarity	Accuracy Precision Objectivity Integrity Correctness Reference accuracy
Completeness/pertinence	Schema Completeness Pertinence **IF5** Instance Value C. Tuple C. Column C. Relation C. Database C.	Completeness (btw different datasets) Pertinence	Completeness	Completeness	Objectivity Completeness

Temporal	Currency IF8 Timeliness, volatility	Recency/temporal accuracy/temporal resolution			
Minimality/redundancy/ compactness/cost	Schema Minimality Redundancy	Redundancy	Minimality		For a law: Conciseness For a legal framework: Minimality, redundancy
Consistency/coherence/ interoperability	Instance Intrarelational Consistency Interrelational Consistency Interoperability	**IF9** Consistency Object consistency Geometric consistency Topological consistency Interoperability	Interoperability	**IF5** Cohesion Referential Temporal Locational Causal Structural **IF5** Coherence Lexical Nonlexical	Coherence Consistency among laws Consistency among legal frameworks
Readability/comprehensibility/ usability/usefulness/ interpretability	Schema **IF7** – Diagrammatic readability Compactness Normalization	Instance Readability/legibility Clarity Aesthetics	**IF5** – Readability, lightness, brightness, uniformityness, sharpness, hue chroma reproductionness usefulness	**IF5** – Readability Comprehensibility **IF5** Cultural readability	**IF6** Clarity Simplicity

Several dimensions in the table are associated with corresponding influencing criteria. Notice:

(a) the great variability of the accuracy cluster with the type of information representation,
(b) the clear distinction between schema and instance related dimensions in the "Structured data" column,
(c) the differentiation in the "Laws and Legal framework" column between qualities of single laws and qualities for the legal framework.

After these general considerations, we discuss more in depth the influence of type of information representation (IF1) on specific dimension clusters listed in the table.

1. *Accuracy* is often considered as an intrinsic IQ dimension (IF9), and its quality level is measured either by comparison with the "true" value (IF5, semantics) or else by comparison with a reference table (IF5, syntax).
2. *Accuracy* for structured data is defined both at the schema level and at the instance level, while for unstructured texts is defined at the instance level, with reference to a weaker property called *structural similarity* (IF4), referring in the word "structural" to the latent internal organization of the text.
3. *Accuracy* for structured data has different metrics for different definition domains. We may focus here on (a) surnames of persons, that are made of one word item (e.g. Smith), or else (b) names of businesses, that may involve several word items (e.g. AT&T Research Labs). When data values are typically composed of one single word, distance metrics are adopted that compare the two words seen as strings of characters, without any further internal structure considered. When data values consist of groups of items, then distance metrics consider the total number of items in data values, and the number of common items (Jaccard's distance), or variants of metrics that are based on the internal structure of values. Even in case of single words, metrics are sensitive to the average length of words in the universe of discourse; so that they change when, e.g., consider surnames in United States and in Asia, where surnames in certain populations are very long.
4. *Spatial accuracy* for maps refers to a bidimensional or tridimensional metric space (IF9).
5. *Consistency* for geographic maps is defined both in the topological space and in the geometric space (IF9).
6. *Cohesion* and *coherence* are proposed for unstructured texts. Both cohesion and coherence represent how words and concepts conveyed in a text are connected on particular levels of language, discourse and world knowledge. Cohesion is considered an objective property (IF6) of the explicit language and text, and is achieved by means of explicit linguistic devices that allow expressing connections (relations) between words, sentences etc. These cohesive devices cue the

reader on how to form a coherent representation. Coherence results from an interaction between text cohesion and the reader. The coherence relations are constructed in the mind of the reader (IF7) and depend on the skills and knowledge that the reader brings to the situation. Coherence is considered a characteristic of the reader's mental representation, and as such is considered subjective (IF6). A particular level of cohesion may lead to a coherent mental representation from one reader but an incoherent representation for another (IF7).

7. *Diagrammatic readability* is usually expressed in terms of the achievement of several aesthetic criteria such as:

   (a) Minimize crossings
   (b) Use only horizontal and vertical lines
   (c) Minimize bends in lines
   (d) Minimize the area of the diagram
   (e) Place most important concept in the middle
   (f) Place parent objects in generalization above child objects.

   Notice that criteria a, b, c and d can be considered syntactic criteria, while e and f are semantic criteria (IF5). Applying such criteria to the two semantically equivalent Entity Relationship diagrams in Fig. 4.7, we may come to the conclusion that the diagram on the right is more readable than the diagram on the left. Unfortunately (or fortunately) this is not a universal conclusion, since about 30 years ago one of the authors was invited to visit Beda University at Peking, and Chinese professors preferred the diagram on the left, claiming that they liked asymmetry and sense of movement (IF7).
8. *Readability of unstructured texts* and *cultural accessibility* refer to the readability/comprehensibility cluster. Readability is usually measured by using a mathematical formula that considers *syntactic features* of a given text, such as complex words and complex sentences, where e.g. complex words are evaluated on the basis of shallow syntax, such as number of syllables. *Cultural readability* refers to difficult (to understand) words, so they are related to the understanding of the word meaning, and as such can be considered more semantic oriented (IF6).
9. Concerning the relationship between IQ dimensions in the different representations vs. objective/subjective measures (IF6), we have produced some figures in the past that confirm the validity of the following intuitive statement in the literature: the less the information is structured, from a restricted domain to a totally unstructured domain, the more subjective measures prevail on objective measures.

In Fig. 4.8 we show two types of information representations, relational tables and diagrams, and three measures of IQ quality, respectively for *accuracy* of data for relational tables, and *readability* for diagrams addressed in previous point 8. It is clear (also recalling the previous example on Chinese professors) that objective measures can be conceived for diagrams, but only to a certain extent, after that we have to deal with human being perceptions.

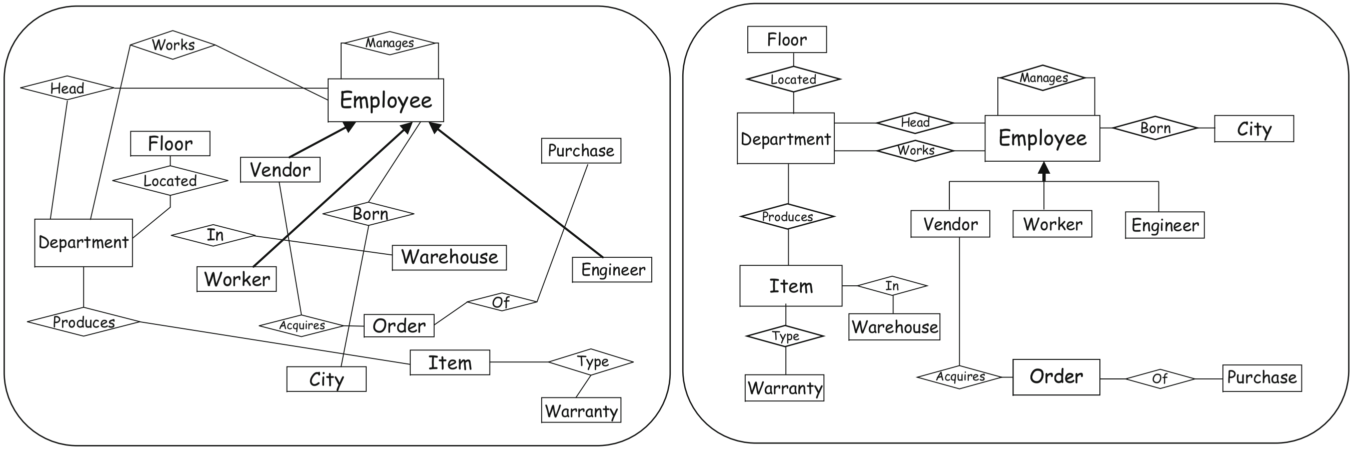
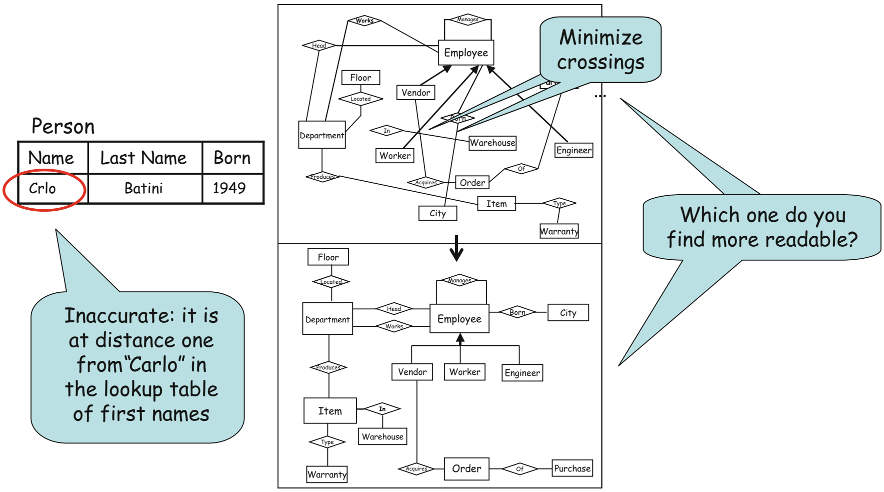


Fig. 4.7 Two semantically equivalent entity relationship diagrams



Person

Name	Last Name	Born
Crlo	Batini	1949

Fig. 4.8 comparison of IQ measures for relational tables and diagrams

4.7 IQ Dimensions and Types of Information Systems (IF3)

We now investigate the relationships between IQ dimensions and the evolution of types of information systems, enabled by the evolution of ICT technologies. The shift from centralized and tightly coupled distributed systems to loosely coupled, distributed and peer to peer systems, and from “controlled” sources to the unrestrainable web results both in bad and in good news from the point of view of IQ. From one side, the overall quality of the information that flows between networked information systems may rapidly degrade over time if both processes and their inputs are not themselves subject to quality control. On the other hand, the same networked information system offers new opportunities for IQ management, including the possibility of selecting sources with better IQ, and of comparing sources for the purpose of error localization and correction, thus facilitating the control and improvement of data quality in the system.

Peer to Peer data management (P2P) Systems, typical of many application areas such as the ones found in the domain of biological databases, differently from centralized and strongly coupled distributed systems do not provide a global schema of the different sources. P2P systems are characterized by their openness, i.e. a peer can dynamically join or leave the system, and by the presence of mappings usually relating pairs of schemas. In P2P systems (and even more in the web) new quality dimensions and issues have to be considered such as *trustworthiness* and *provenance*.

The evaluation of the trustworthiness (or confidence) of the data provided by a single peer is crucial because each source can in principle influence the final, integrated result. A common distinction is between the reputation of a source, which refers to the source as a whole, and the trust of provided data, e.g., the trust of the mapping that the source establishes with the other sources in a P2P system. While several trust and reputation systems have been proposed in the literature (see Josang et al. (2007) for a survey), there is still the need to characterize the trust of a peer with respect to provided data and use such information in the query processing step. Effective methods for evaluating trust and reputation are needed, with the specific aim of supporting decisions to be taken on result selection.

Information provenance describes how data is generated and evolves with time going on, which has many applications, including evaluation of quality, audit trail, replication recipes, citations, etc. Generally, the provenance could be recorded among multiple sources, or just within a single source. In other words, the derivation history of information could take place either at schema level (when defined), or at instance level. Even if significant research has been conducted, a lot of problems are still open. For the schema level, the most important are query rewriting and schema mappings including data provenance, and for the instance level, we mention relational data provenance, XML data provenance, streaming data provenance (Buneman and Tan 2007). Moreover another important aspect to be investigated is dealing with uncertain information provenance for tracking the derivation of information and uncertainty.

4.8 IQ Dimensions and Levels of Semantic Constraints (IF4)

Influencing factor IF4 deserves special attention in the context of this book. We address in this section the discussion on levels of semantic constraints and the adoption of OWA vs. CWA, while next section details the changes in perspective when moving form databases to ontologies and knowledge bases.

As we anticipated in Sect. 4.3, different levels of semantic constraints can be imposed to data. In databases, data and schemas are tightly coupled; schemas pre-exist to data and control methods implemented by database management systems can enforce data to comply to the schema, which, even if poorly, defines their semantics. As an example, normal forms in relational databases are defined at the schema level, and are expressed in terms of properties of functional dependencies defined in relational schemas. A relational database whose relation schemas are in normal form, has relation instances free of redundancies and inconsistencies in updates, since every "fact" is represented only once in the database.

The coupling of data and schemas in semi-structured or weakly structured data is way looser. Even when languages for semi-structured or weakly structured data are accompanied with languages for describing data schemas, e.g., XML-Schema for XML, RDFS and OWL2 for RDF (Antoniou and van Harmelen 2008), schemas are not required to pre-exist to data and the enforcement of the compliance of data to a schema at publishing time is weaker (it is left to the data publisher). Data in these cases are associated with schemas by means of annotation mechanisms. Finally, the use of metadata, e.g., based on folksonomies, or other annotation schemes, can be seen as a way to associate data with schema-level information that provides data with semantics. However, the maximum freedom achieved by these representation approaches leads to a yet weaker coupling of data and schemas.

As an example, let us focus on semantic data represented in RDF, which is also accompanied with expressive languages for the representation of schemas. A schema for RDF data can be defined by a RDFS vocabulary; however, there is no mechanism to enforce data to be compliant to the schema; even using reasoning, RDFS is not expressive enough to detect inconsistencies, because of its deductive semantics (the schema is used to make inference, not to constraint their meaning) and the lack of expressivity (concept disjointness and cardinality restrictions cannot be modeled in RDFS) (Antoniou and van Harmelen 2008); although counterintuitive inferences can be considered a measure of poor compliance between data and schemas (Yu and Heflin 2011), no inconsistencies can be detected, making a quality dimension such as *soundness* difficult to assess.

In addition, the adoption of CWA or OWA has an influence on this discussion; OWA has an impact on the difficulty of defining and evaluating the compliance between data and schemas: a relation between two instances can hold even if the schema does not model such relation between the concepts the instances belong to; conversely, we cannot conclude that a relation between two concepts of different schemas does not hold because it is not represented in the data instances.

4.9 The Impact of the Information Representation Model Flexibility on IQ

4.9.1 From Databases to Knowledge Bases on the Web

Considering the remarks in the previous section, we can conclude that the more types of information are considered, and the more diverse and decentralized information management models and architectures are, the more we are in need of rethinking the perspective through which we look at information quality (in computer science). An interesting perspective on the role that diversity of information objects can play in IQ emerges if we investigate how the IQ perspective changes when moving from data bases to web knowledge bases (KBs), i.e., knowledge bases published, shared and accessible on the web. Web KBs are, in fact, diverse and often decentralized information sources.

In general, we can see a web KB as composed of a terminological component and an assertional component (Staab and Studer 2004). The terminological component of a web KB, usually called *ontology*, conveys general knowledge about a domain in terms of logical constraints that define the meaning of the concepts (and relations) used in the language (e.g. "*every Cat is an Animal*"); ontologies for web KBs (web ontologies for short) are represented with web-oriented formal languages like OWL, RDFS, and SKOS (Antoniou and van Harmelen 2008).

The assertional component of a web KB expresses facts in terms of properties of individuals, i.e., instances of ontology concepts, and relations holding between them (e.g. "*Fritz is a Black Cat*"; "*Fritz is friend of Joe*"). We remark that the distinction between the two components in a KB can be more or less sharp depending on the language used to represent the KB and the ontology, but it can be adopted without loss of generality for our purposes.[1] Also, terminological and assertional components can be independent (see the Sect. 4.9.2) and several ontologies that are not designed for specific assertional components exist, e.g., consider an upper-level ontology such as DOLCE.[2]

In the following we focus on IQ as investigated in the field of ontologies because they represent a fundamental aspect of web KBs.[3]

4.9.2 Some Issues Arising from the Investigation of IQ for Ontologies: Semiotic, Diversity, Reuse

We concentrate on three main characteristics of ontologies, each of which shed light on significant aspects of IQ when considered in an open information spaces.

4.9.2.1 Ontologies Are Semiotic Objects

One of the first works that addressed the problem of evaluating (the quality of) ontologies exploited a framework based on a semiotic model (Burton-Jones et al. 2005). A similar approach appears in a model that describes the relationship between ontologies as formal (externalized) specifications, (mental) conceptualization and the "real world" (Gangemi et al. 2006). Within this cognitive-flavored semiotic approach, several quality dimensions, and metrics have been defined on top of these frameworks. Gangemi et al. (2006) distinguishes between quality dimensions and evaluation principles.

Three types of dimensions under which it is possible to evaluate an ontology are discussed. The *structural dimension* focuses on syntax and formal semantics, i.e. on ontologies represented as graphs (context free metrics). The *functional dimension*

[1] The use of lexical resources such as WordNet or other taxonomies represented in SKOS in KBs is widespread. Although these resources are used for annotation purposes in the assertional components of KBs, they are very often referred to as *ontologies* in the community (Manaf et al. 2012) and share likewise terminological components of KBs define semantic relations between concepts in a domain.

[2] http://www.loa.istc.cnr.it/DOLCE.html

[3] Most of these approaches explicitly consider ontologies as KB terminologies represented in web-compliant formal languages. Some of the approaches use a even broader definition of ontology which includes instances and relations among instances and is equivalent to our definition of web KB.

is related to the intended use of a given ontology and of its components, i.e. their function in a context. The focus is on the conceptualization specified by an ontology. The *usability-profiling dimension* focuses on the ontology profile (annotations), which typically addresses the communication context of an ontology (i.e. its pragmatics). Then several principles (or evaluation-driven dimensions) are introduced, namely: *cognitive ergonomics, transparency, computational integrity and efficiency, meta-level integrity, flexibility, compliance to expertise, compliance to procedures for extension, integration, adaptation, generic accessibility*, and *organizational fitness*.

Following the cognitive flavor of this point of view, a quite recent approach studied a measure of cognitive quality based on the adequacy of represented concept hierarchies w.r.t. the mental distribution of concepts into hierarchies according to a cognitive study (Evermann and Fang 2010). These cognitive approaches clarify an important issue that has been central in the research about IQ in the ontology domain: ontologies are knowledge objects that are used by someone and for some specific goals; the evaluation of the quality of ontology should consider ontology in its semiotic context.

4.9.2.2 Ontologies as Diverse Knowledge Objects

As it can be captured from the broad definition of ontology given at the beginning of this paragraph, ontologies are very different one from another. Some ontologies are flat, while some others consist in deep concept hierarchies; some ontologies are deeply axiomatized, while others, e.g. Geonames,[4] look more like database schemas (Cruz et al. 2012, 2013). Moreover, often ontologies cannot be modified but are reused and eventually extended. Some metrics defined for evaluating an ontology can be adopted to provide a value judgment about an ontology. Other metrics proposed so far are more intended as analytic dimensions to profile an ontology, and to understand its structure and its properties. As an example, one of the first unifying framework proposed to assess ontology quality distinguishes between syntactic, semantic, pragmatic and social qualities (see Table 4.5) (Burton-Jones et al. 2005).

Although lawfulness and interpretability clearly lead to a value judgment (positive vs. negative), metrics such as richness and history can be hard to be associated with a value judgment. In other frameworks such as the one proposed by (Gangemi et al. 2006; Tartir et al. 2005), which put a lot of focus on the computability of the defined metrics, most of the metrics are more aimed at profiling an ontology, rather than at assessing its quality from a value perspective. The idea is that these quality metrics can be used to summarize the main property of an ontology and their evaluation can be used by third party applications. As an example, a machine learning method that takes advantage of fine-grained ontology profiling techniques (extended from Tartir et al. (2005)) to automatically configure an ontology matching system

[4] http://www.geonames.org/

Table 4.5 Types of qualities and dimensions in Burton-Jones et al. (2005)

Dimension	Metrics	Definition
Syntactic quality	Lawfulness	Correctness of syntax
	Richness	Breadth of syntax used
Semantic quality	Interpretability	Meaningfulness of terms
	Consistency	Consistency of meaning of terms
	Clarity	Average number of word senses
Pragmatic quality	Comprehensiveness	Number of classes and properties
	Accuracy	Accuracy of information
	Relevance	Relevance of information for a task
Social quality	Authority	Extent to which other ontologies rely on it
	History	Number of times the ontology has been used

has been recently proposed (Cruz et al. 2012). These approaches, which consider ontologies also as computational resources (see point above), differ from early works on ontology quality that were based on philosophical (metaphysical) principle to establish the quality of an ontology as a conceptual model, but whose analytical principles are more difficult to be made computable.

4.9.2.3 Ontologies as (Reusable) Computational Resources

A key aspect of ontologies is that they are expected to be reused by other ontologies, applications, or, more generically, third party processes. It is often the case that one has to select an ontology to reuse it in a given domain. Ontologies can be used to support search or navigation. Different aspects of an ontology can be more or less amenable depending on the task an ontology is aimed to support. Approaches that evaluate ontologies on a task basis (Yu et al. 2007; Lei et al. 2007; Strasunskas et al. 2008) seem to have received more attention, recently, than previous approach based on metaphysical and philosophical considerations (Guarino and Welty 2002), which better fit the use of ontologies as conceptual models, rather than as computational objects.

4.10 Conclusive Remarks

In this paper we have discussed the main issues considered in data quality and information quality research, identifying several factors influencing them. According to a quite common use of the terms in the technical literature published by the data management community, we referred to data quality when structured data where addressed, and to information quality when information represented according to other data models is considered. However, the consideration of information digitally represented by different types of data and organized according to different data models has definitely a deep impact on the most relevant issues considered in information quality, including the definition itself. The more heterogeneous the considered information is, the more a comprehensive theoretical framework defining in a general way the mutual relationship between several crucial concepts in the definition and assessment of information quality (e.g., data, information, information carrier, observer, task, and so on) is needed. Recent works in the field of ontology evaluation framed the (information) quality problem within a broader semiotic and cognitive framework (see Gangemi et al. (2006) and Evermann and Fang (2010)). A similar concern can be found in several works on information quality coming from the Information Systems community (see Wand and Weber (1995, 1990) and Wand and Wang (1996)). These approaches can provide important contributions to a theoretical clarification of the common use of information quality core concepts and issues, in a context where the amount and the degree of complexity, diversity, and interconnection of the information managed in ICT is constantly increasing.

One problem that we believe particularly interesting is tightly related to the influencing factor IF4 addressed in this paper, which considers the impact on information quality of the degree of coupling between data and schemas (where available), and the difference in the semantics associated with structured and other types of data (e.g., schemaless data such as texts, images, sounds). An interesting research question concerns the extent to which information quality is affected by the degree of coupling between data and schemas, or, more in general, the role played by semantics defined by data models and schemas in the definition of information quality. This issue tightly relates to the relationship between data, information and *truth* in information systems. In this case, information quality faces the dualism of scheme and content, of organizing systems and something waiting to be organized, as criticized by Davidson as the third dogma of empiricism (Davidson 1974). If schema-driven data can be easily interpreted as carriers of factual information and interpreted according to a semantic theory of truth (Kirkham 1992) (e.g., through mapping to First-Order Logic), the connection between other types of information representations (e.g., maps, images, sounds) and factual information has been less investigated and results more obscure. Texts can be taken as borderline examples from this point of view: most of textual documents are clearly carriers of factual information to a human reader, but their digital representation is by no means related to any factual interpretation (hence, investigations in the field of natural language processing, knowledge extraction, and so on).

Moving to the conceptual challenges to be faced in the future, as also shown by the above reference to the work of Davidson, it is our point that contributions from philosophy can bring some theoretical clarification to IQ basic issues and influencing factors. Otherwise, we argue that these challenges are going to be tangled by dichotomies such as the ones implied in the discussion carried out in previous sections. As an example, consider factual information, which is represented both in structured and semi-structured information. Some of the quality dimensions proposed in the literature pose the question of adherence of a certain representation to real world (see for example IF6, and BI2 as for clusters of dimensions such as accuracy or consistency). As for these issues, considering (IF4), the critical question here is whether information qualities pertain to facts of sense or rather to laws of logic or, else, whether IQ is a matter of synthetic rather than analytic knowledge (e.g., are there truly intrinsic quality dimensions?). This and other issues related to IQ and discussed in the paper recall in fact philosophical disputes about the two dogmas of empiricism, against which Quine provided arguments, in favor of a holistic perspective. On the one hand, Quine rejected the distinction between truths independent from facts, and truths grounded in facts; on the other hand, he contrasted reductionism as the theory according to which the meanings of statements come from some logical construction of terms, exclusively referring to immediate experience (Quine 1951).

An example is the current debate among scholars and practitioners about the use of quality dimensions coming from practice in specific domains (Embury et al. 2009), instead of well-established (often academic) ones. Furthermore considering practitioners' debate on Linkedin Groups (e.g., the IAIDQ – Information/Data Quality Professional Open Community) where some members argue the need for better definition of "data quality" as different from "data qualities",[5] and of "dimensions",[6] likewise. As for a lesson learned by the work of Quine, this issue may require challenging the ontology anchoring data quality since (Wand and Wang 1996). In particular, we believe that the following assumptions are worth being challenged when conducting research on information quality "in the wild":

- *the quality of the data generated by an information system depends on the design of the system*: this assumption is grounded in a closed perspective on the information system design as bound by an organization requirements; whereas today we assist to an ecosystems of information systems, providing information to both businesses and lay users, often in an open and bidirectional way, actually having different design requirements for intended use by different organizations and target users.
- *The internal/external views of an information system*: strictly related to the previous assumption, this dichotomy leads to a focus on the internal view considered as use-independent, and the identification of a set data quality dimensions comparable across applications and viewed as intrinsic to data. This perspective is based on the idea that systems requirements capture the true intentions of the users (Wand and Wang 1996) As said above, today it is difficult to identify the true intentions of the users, due to the variety, heterogeneity, and the openness of the information systems, thus questioning the internal view assumption: "issues related to the external view such as why the data are needed and how they a used are not part of the model" (Wand and Wang 1996, p. 11).
- *The definition of data deficiency as inconformity* between a view of a real-world system mediated by a representing information system, and a reference view of

[5] See IAIDQ discussion "Do data quality dimensions have a place in assessing data quality?", 2nd July 2013.

[6] See IAIDQ discussion "Do data quality dimensions have a place in assessing data quality?", 9th July 2013.

a real-world system obtained by direct observation. Again, today information systems are not designed "from scratch" and are composed both by legacy systems and a (often) dynamic configuration of external systems for information search, retrieval, and production (social networks, internet of things, etc.). Thus, *inconformity* between views of real-world is actually difficult to ascertain, being today probably a rule rather than an anomaly of information systems "in the wild" (as for this issue, the arguments by Quine on the indeterminacy of translation and the meaning of the expressions of one's own language (Weir 2008; Quine 1960) may provide insights to information quality research).

Furthermore, the above issues may also be related to the problem of knowledge of things *by acquaintance* (e.g. in the case of images) and *by description* (e.g. in the case of structured data), as stated for example by Bertrand Russell: "We shall say that we have acquaintance with anything of which we are directly aware, without the intermediary of any process of inference or any knowledge of truths" (Russell et al. 1910). Thus, differently from knowledge by acquaintance, knowledge by description connects the truths (carried by data, in our case) with things with which we have acquaintance through our direct experience with the world (*sense-data*, in the Russell perspective).

As an example of the role of factual information carried by data in information quality, consider the above discussion pointing out data and information quality pose the question of adherence of a certain representation to real world (see for example, clusters of dimensions such as *Accuracy/correctness/precision* or *Completeness/pertinence*). This question points to one of the most controversial issues discussed in philosophy so far. Significantly, Russell discusses this issue using the term *data*, and in particularly distinguishing between *hard data* and *soft data:* "The hardest of hard data are of two sorts: the particular facts of sense, and the general truths of logic" (Russell 1914/2009, p. 56).

Indeed, from the above discussion we could ask ourselves to which extent information quality (and specific quality dimensions) may pertain to the domain of both hard and soft data. Thus, the critical question is if information quality pertains to facts of sense or rather to laws of logic, which play a fundamental role both at the data model level (e.g., relational algebra for relational databases) and at the schema level (e.g., all persons are identified by their Social Security Number). Again, what can we say about data that are not straightforwardly associated with any truth-based semantics (e.g. images)? Finally, we mention that the role of the processes and tasks that are supported by an information system has to be considered when investigating the above research questions (the number of papers focusing on task-oriented evaluation of information quality is in fact increasing, see, e.g. Yu et al. (2007), Lei et al. (2007), Strasunskas et al. (2008)).

We believe that the above insights should be considered working constructs, with the aim of investigating whether philosophical research can help to clarify significant relationships between basic issues and influencing factors of IQ, too often narrowly considered under a technical perspective in computer science and information systems areas. In particular, the previously cited well known contributions from philosophy may help bending IQ basic issues and influencing factors towards a holistic or else pragmatist perspective (Rorty 1982); this latter being suitable to challenge what we have described in the introduction as the current wild landscape in which information is published, processed and consumed.

Acknowledgments We acknowledge Raimondo Schettini and his research group for providing insights and some of the figures in the paper, with specific reference to image quality.

References

Antoniou, G., & van Harmelen, F. (2008). *A semantic web primer*. Cambridge: MIT Press.

Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques*. Berlin/Heidelberg: Springer.

Batini, C., Di Battista, G., & Santucci, G. (1993). Structuring primitives for a dictionary of entity relationship data schemas. *IEEE Transactions on Software Engineering*, *19*, 344–365.

Batini, C., Cabitza, F., Pasi, G., & Schettini, R. (2008). *Quality of data, textual information and images: A comparative survey*. Tutorial at the 27th International Conference on Conceptual Modeling (ER 2008), Barcelona, available on request to batini@disco.unimib.it.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*, 16:1–16:52.

Batini, C., Palmonari, M., & Viscusi, G. (2012, July 2–6). The many faces of information and their impact on information quality. In P. Illari & L. Floridi (Eds.), *Information quality symposium at AISB/IACAP World Congress*, Birmingham (pp. 5–25). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Buneman, P., & Tan, W. (2007). *Provenance in databases*. SIGMOD Conference (pp. 1171–1173), ACM Press.

Burton-Jones, A., Storey, V. C., Sugumaran, V., & Ahluwalia, P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering, 55*, 84–102.

Ciocca, G., Marini, F., & Schettini, R. (2009). Image quality assessment in multimedia applications. *Multimedia Content Access Algorithms and Systems III, SPIE*, Vol. 7255, 72550A.

Cruz, I. F., Fabiani, A., Caimi, F., Stroe, C., & Palmonari, M. (2012). Automatic configuration selection using ontology matching task profiling. In *ESWC 2012* (pp. 179–194).

Cruz, I. F., Palmonari, M., Caimi, F., & Stroe, C. (2013). Building linked ontologies with high precision using subclass mapping discovery. *Artificial Intelligence Review, 40*(2), 127–145.

Davidson, D. (1974). On the very idea of a conceptual scheme. In J. Rajchman & C. West (Eds.), *Proceedings and addresses of the American Philosophical Association*, Vol. 47 (1973–1974), pp. 5–20. JSTOR.

Embury, S. M., Missier, P., Sampaio, S., Greenwood, R. M., & Preece, A. D. (2009). Incorporating domain-specific information quality constraints into database queries. *Journal of Data and Information Quality, 1*, 11:1–11:31.

Encyclopedia of GIS. (2010). *Encyclopedia of geographical information systems*. Springer.

Evermann, J., & Fang, J. (2010). Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems, 35*, 391–403.

Floridi, L. (2011). Semantic conceptions of information. The Stanford Encyclopedia of Philosophy.

Gangemi, A., Catenacci, C., Ciaramita, M., & Lehmann, J. (2006). Modelling ontology evaluation and validation. In Y. Sure & J. Do-mingue (Eds.), *ESWC* (pp. 140–154), Vol. 4011 of Lecture Notes in Computer Science, Springer.

Gasparini, F., Marini, F., Schettini, R., & Guarnera, M. (2012). A no-reference metric for demosaicing artifacts that fits psycho-visual experiments. *EURASIP Journal on Advances in Signal Processing, 2012*, 4868–4873.

Guarino, N., & Welty, C. A. (2002). Evaluating ontological decisions with OntoClean. *Communications of the ACM, 45*, 61–65.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge: MIT Press.

ISO/IEC FDIS 25012. (2008). Software engineering – Software product quality requirements and evaluation – Data quality model.

Josang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems, 43*, 618–644.

Kirkham, R. (1992). *Theories of truth: A critical introduction* (pp. xi, 401). Cambridge, MA: The MIT Press.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information and Management, 40*, 133–146.

Lei, Y., Uren, V. S., & Motta, E. (2007). A framework for evaluating semantic metadata. In D. H. Sleeman & K. Barker (Eds.), *K-CAP* (pp. 135–142). ACM.

Lindland, O. I., Sindre, G., & Solvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE Software*, *11*, 42–49.

Liu, L., & Chi, L. (2002). Evolutional data quality: A theory-specific view. In *The 6th International Conference on Information quality*, Boston.

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality, 1*, 1–22.

Manaf, N. A. A., Bechhofer, S., & Stevens, R. (2012). The current state of SKOS vocabularies on the web. In *ESWC 2012* (pp. 270–284). Berlin/Heidelberg: Springer-Verlag.

Merriam Webster. Knowledge. http://www.merriam-webster.com/dictionary/knowledge

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60*, 20–43.

Quine, W. V. O. (1960). *Word and object*. Cambridge: MIT Press.

Rorty, R. (1982). *Consequences of pragmatism: Essays, 1972–1980*. Minneapolis: University of Minnesota Press.

Russell, B. (1910). Knowledge by acquaintance and knowledge by description. In *Proceedings of the Aristotelian Society* (New Series), Vol. XI (1910–11), pp. 108–128.

Russell, B. (1914/2009). *Our knowledge of the external world*. London/New York: Routledge.

Staab, S., & Studer, R. (Eds.). (2004). *Handbook on ontologies*. Berlin: Springer.

Strasunskas, D., & Tomassen, S. L. (2008). Empirical insights on a value of ontology quality in ontology-driven web search. In R. Meersman & Z. Tari (Eds.), *OTM Conferences (2)*, Vol. 5332 of Lecture Notes in Computer Science (pp. 1319–1337). Springer.

Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., & Aleman-Meza, B. (2005). *OntoQA: Metric-based ontology quality analysis*. IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources.

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

Wand, Y., & Weber, R. (1990). An ontological model of an information system. *IEEE Transactions on Software Engineering, 16*, 1282–1292.

Wand, Y., & Weber, R. (1995). On the deep structure of information systems. *Information Systems Journal, 5*, 203–223.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*, 5–33.

Weir, A. (2008). Indeterminacy of translation. In E. Lepore & B. C. Smith (Eds.), *The Oxford handbook of philosophy of language* (pp. 233–249). Oxford: Oxford University Press.

Yu, Y., & Heflin, J. (2011). Extending functional dependency to detect abnormal data in RDF graphs. In L. Aroyo, C. Welty, H. Alani, J. Taylor, & A. Bernstein (Eds.), *The 10th International Conference on The semantic web – Volume Part I (ISWC'11)* (pp. 794–809). Berlin/Heidelberg: Springer Verlag.

Yu, J., Thom, J. A., & Tam, A. (2007). Ontology evaluation using Wikipedia categories for browsing. In *The Sixteenth ACM Conference on Information and knowledge management*, *CIKM '07* (pp. 223–232). New York: ACM.