

Toward a Unified Model for Information Quality

C. Batini, D. Barone, F. Cabitza,
G. Ciocca, F. Marini, G. Pasi and R. Schettini
Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli studi di Milano-Bicocca
20126, Viale Sarca 336
Milano, Italy

{batini,barone,cabitza,ciocca,marini,pasi,schettini}@disco.unimib.it

ABSTRACT

We present a model which allows to define in an uniform way information quality dimensions related to heterogeneous types of information, such as structured data managed in data bases, semi-structured and unstructured texts and images. We first define a set of concepts that allow to represent several basic characteristics of such heterogeneous types of information. Then, we introduce a general categorization of quality dimensions and sub-dimensions, which are then specialized to structured data, semi- and unstructured texts and images. In so doing, we provide, to our knowledge, a first attempt to unify the information quality issue for heterogeneous information types.

Keywords

data quality, information quality, structured data, semistructured data, unstructured text, image, quality dimension, quality metrics.

1. INTRODUCTION

In last twenty years the concept of quality in organizations has been declined in several ways both in private consultancy and academic research domains. Quality in an organization relates to the ability of the organization to fulfill the needs and expectations of its customers and users, to create value in an efficient and effective manner, to control and improve the performance of its processes and the outcomes of its services. With the advent of office automation technologies, and of the digitalization of information systems, organizational quality has increasingly been bound and coupled with the quality of the data that an organization produces, holds and consumes to retain and retrieve valuable information. Now more than ever, information is available in different formats, media and resources and it is accessed and exploited through multiple channels. Available information is not only textual, it is also represented by pictures, videos and sound, all contributing to create the in-

formation assets of an organization. Since all these kinds of information are intertwined and can refer to the same conceptual entities in complex and complementary ways, the assessment of information quality should take all of them into account, with appropriate dimensions and metrics. In this paper, we address the problem of assessing information quality when information can be carried by either structured data, semi-structured and unstructured texts, or images. To this aim a unified model for information quality is proposed.

The paper is organized as follows. In Section 2 we define the basic concepts on which our model is grounded; in Section 3 we define the quality model. Based on this approach, and due to the multidimensional nature of information quality, in Sections 4, 5 and 6 we address each type of considered information content in a separate way, i.e. structured alphanumeric data, semi- and unstructured texts and images. Section 7 concludes the paper.

2. INFORMATION QUALITY

Formally, we refer to a *digital information item* as to an atomic information unit in a considered context. A *digital information item* is constituted by a *content*, a *carrier*, and possibly (this is not mandatory) by a *schema* (see Figure 1). A set of digital information items constitutes a *digital information resource*, on which users can rely to get some information. Any digital information resource must be represented in some way, by what we call a *digital information representation*. Common examples of digital information representation are databases, inverted indexes and raster pictures. The *content* of a digital information item is constituted by an *essence* and possibly by some *metadata* and *annotations* associated with it. *Essence* is the real material itself, the result of the creative process, what can be used to produce value or to inform actors and their processes. *Metadata* may be associated with the essence in order to characterize a set of its properties (like the name of the author or the date of creation). *Annotations* constitute any additional and explanatory information that may be associated with the essence. The *content* is then simply the combination of essence, metadata and annotations, i.e., the kernel of information plus what describes and enriches it. While the essence must exist for a content to exist, metadata and annotations can also be absent. The essence of a *content* can be of different types: in this paper, we consider that essence may be expressed in terms of i) *alphanumeric data*, namely symbols, numbers and terms pertaining to a specific domain of values; ii) *texts*, namely self-explanatory sentences, narra-

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permission from the publisher, ACM.

VLDB '08, August 24-30, 2008, Auckland, New Zealand
Copyright 2008 VLDB Endowment, ACM 000-0-00000-000-0/00/00.

tive passages, or any written work; and iii) images, namely, visual content which encodes information either about the geometry of scenes and the properties of the objects located within these scenes [15], or about a visual representation of a non-visual information, like in the case of maps. Alphanumeric and textual content can be further classified with respect to its *level of structuredness*, i.e., as *structured content*, *semi-structured content* and *unstructured content*. When considering alphanumeric content, this level of structuredness varies according to the *schema*, e.g., whether data types, domain types and semantic constraints are explicitly expressed or not. When considering texts, the level of structuredness varies according to whether a *schema* exists (as in the case of XML-marked up semi-structured text) or whether a reader can recognize any meaningful structure within the *carrier*.

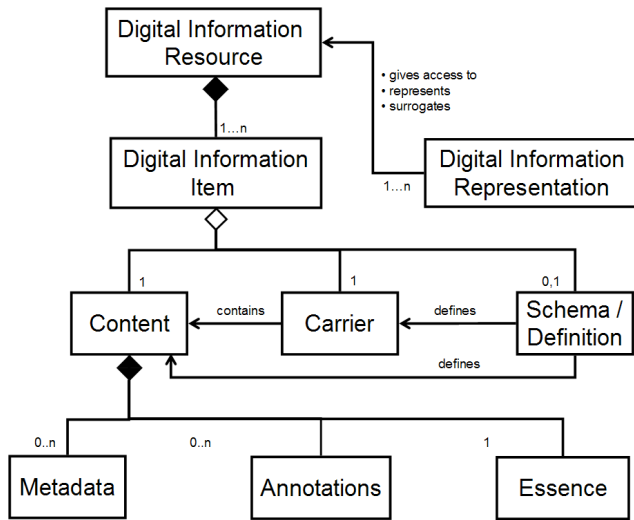


Figure 1: The basic concepts of the paper and their main relationships.

With the term *schema* we denote the way in which the structure of the content and its semantics are explicitly declared in terms of formats, types, constraints and relationships between either the values (e.g., in structured data), fields (e.g., in structured and semi-structured text), sections (e.g., in unstructured but formatted text) or partitions (e.g., in images) of the digital content. The concept of *carrier* of the content of a digital information item is strongly related with the concept of schema.

We introduce an example to clarify the concepts previously defined. Let us consider a digital information item constituted by an XML file (document). The schema of such an information item is constituted by the XML schema. An example of content is the name ‘Valentino Rossi’ embedded in a tag defining names as alphabetic strings of max 25 characters. In this case the carrier is the tag ‘name’ defined in the XML file and whose inner value is ‘Valentino Rossi’. Metadata and annotations are not mandatory. An example of metadata would be the ‘creation date’ and ‘file dimension’ of the above mentioned XML file. An annotation would be any external resource that, after the creation of the XML file, is linked to the XML file for later reference and addition. Readers can also refer to Figure 2 for more examples.

From the above general perspective, any digital information item can be seen as a content, whose characteristics change depending on either the context of use, the representational medium, or the goal of use. In this context, by *Information Quality* (IQ) we refer to the capability of a set of digital information items to meet a set of requirements, either explicitly or implicitly expressed, pertaining to their use in a given context under specific conditions. The quality of a digital information item can be assessed with respect to its content, and according to a specific schema, if any. This means that quality can be assessed with respect to both the essence of an information item, as well as to its metadata and annotations. Hence, the overall quality of a digital information item can be the composition of assessments carried out for different types of content. Moreover, even with respect to content, IQ can be assessed in different ways according to the type of content.

Types of Items	Schema	Content	
		Metadata Annotations	Essence
Structured Alphanumeric Data (e.g. Data bases)	<ul style="list-style-type: none"> • Conceptual schema • Logical schema (conceptual + attributes + constraints, etc.) 	<ul style="list-style-type: none"> • Properties (e.g., author, owner, creation date, table dimension) • Access (read/write) rights. 	<ul style="list-style-type: none"> • Values, i.e., data items in single rows of a table.
Semi-Structured Text (e.g., XML Documents with a Document Definition)	<ul style="list-style-type: none"> • XML Schema • Document Type Definition 	<ul style="list-style-type: none"> • Properties (e.g., author, owner, creation date, file dimension). 	<ul style="list-style-type: none"> • Values, text items within specific tags.
Unstructured but well-formatted Text (e.g., research papers)	<ul style="list-style-type: none"> • Document Structure Description (e.g., mandatory items, section hierarchy, max length) 	<ul style="list-style-type: none"> • Properties (e.g., authors, creation date, keywords). • Any inscription put “in the margin” 	<ul style="list-style-type: none"> • Sections of written text with specific formatting (font dimension and style).
Unstructured, unformatted text (e.g., a narrative passage)	<ul style="list-style-type: none"> • Generally implicit or absent (e.g., the general plan, the outline, the boilerplate) 	<ul style="list-style-type: none"> • Properties (e.g., authors, creation date, keywords). • Any inscription put “in the margin” 	<ul style="list-style-type: none"> • Set of textual sentences.
Digital Images (e.g., an X-Ray picture, a camera photo, a geographic map)	<ul style="list-style-type: none"> • Generally implicit or absent (e.g., geometriotopological schema of human faces in passport photos) 	<ul style="list-style-type: none"> • Properties (e.g., authors, creation date, resolution) • Any inscription justaposed in a specific layer “over” the picture. 	<ul style="list-style-type: none"> • Set of pixels (picture elements).

Figure 2: Examples of schema and content in the digital domain.

In this paper, we aim to contribute to the development of a comprehensive approach to IQ modeling of digital information items such as multimedia documents, i.e. documents composed by alphanumeric structured data, text, and images. Hence we aim to define a general model of information quality that can be used to conveniently assess and improve the quality of information produced, used and exchanged in any organizational domain. The proposed model can help analysts and users in keeping track of the quality factors that are of interest in the domain of digital information.

In what follows, more specific definitions will be proposed for each content type in order to characterize the related quality dimensions. To our knowledge, our contribution is a pioneering one, although similar approaches are being carried out by standardizing bodies world-wide. Accordingly, we refer to the comprehensive Data Quality Model that is under development within the ISO project SQuaRE (Software Product Quality Requirements and Evaluation). At the present moment, this model is denoted as ISO/IEC FCD 25012.2 standard. Currently, the development of the ISO/IEC 25012 standard is at the enquiry stage (i.e., its current Draft International Standards is on ballot). In what follows, we will briefly denote this standard as the ‘ISO standard’.

In the research literature, IQ is modelled in terms of what are usually called *dimensions* or *characteristics*. All IQ dimensions can be traced back to three macro-categories: *in-*

trinsic, external and contextual [19]. These approximately correspond to the internal, external and in-use categories of the ISO standard. Intrinsic dimensions depend on the information itself and its schema (if any), irrespectively of where and how it is used. External dimensions depend on the technological environment and tools where/by which the information is used, and on the technological properties of the information system that encompasses it (either at the software or hardware level). Contextual dimensions depend on the actual environmental, organizational and socio-technical context where information is accessed, used and produced.

Dimensions and sub-dimensions related to intrinsic quality provide criteria upon which to guarantee, assess and improve the quality with respect to either the values (i.e., essence) and the schema (if it exists). Aspects related to the external quality regard properties of the components of the system that manages and provides access to information (e.g., the software that indexes and retrieves unstructured text). Contextual dimensions regard the capability of information to enable users to achieve their own objectives, and to add values into their activity in a specific context. Intrinsic and external aspects are usually measured by means of quantitative and objective metrics. Contextual dimensions are usually expressed in terms of the subjective point of view of actual users by their degree of satisfaction with respect to how data fulfill their operative needs.

3. INFORMATION QUALITY DIMENSIONS, SUB-DIMENSIONS AND METRICS

In this section, we present the basic, well established approach we adopt for defining a unified model for the assessment of IQ of digital multimedia information items. Based on this approach, and due to the multidimensional nature of IQ, in Sections 4, 5 and 6 we address each type of considered information content in a separate way, although complying to the same basic model.

As mentioned in Section 2, *quality dimensions* identify significant and possibly inter-dependent aspects concurring to the assessment of information quality. Based on the well established approach to the definition of quality dimensions in the context of IQ, the attributes of digital information items are categorized into main *IQ dimensions*, which are further classified into *IQ sub-dimensions*, in their turn decomposed in lower-level and measurable *IQ metrics*.

We introduce three high-level dimensions, namely *soundness, usability* and *portability*; these are called high-level dimensions, to distinguish them from quality sub-dimensions that refine the high-level ones.

Soundness is the characteristic of being free from defects, good, accurate, complete and reliable. This dimension is related to the extent a content can satisfy either some stated or implicit requirements when used under specified conditions. Soundness can be measured either subjectively or objectively, by either comparing the content to some “gold standard” (e.g., the original version of a text, a target image), which is assumed to be perfect and as sound as possible, or by considering no reference but by taking into account a *correlating model* (e.g., a perceptual model of human vision or a set of grammatical and orthographic rules) between a specific sub-dimension and the value of soundness. Soundness can be seen either from an intrinsic perspective, when the requirements are independent of any technological

system; or from an external point of view, when the specified conditions of use involve using some actual technological system.

Usability is related to the characteristic of being of some utility to some user. In other words, this dimension is related to the extent a content can satisfy the actual needs of either who (user) or what (process) exploits it. Usability depends both on the system used to either reproduce or access the content, and on the purpose underlying its use, reproduction and access. Usability is related then to the semantic aspects of a content, i.e., what it means, what objects of interest it represents for a specific aim. Usability encompasses characteristics that pertain to both external and contextual aspects. That is, this is a dimension the value of which depends on the technological layer that makes data usable, as well as on the context of production/use and re-use of the content. Accordingly, usability is usually measured in terms of subjective and user-centered assessments, although also objective measures could be related to usability under simple assumptions (e.g., the more a content is used, in terms of number of accesses and time of use, the more it is usable).

Portability is related to the extent the content is useful also outside the original context of production and use, i.e., the extent to which the content can satisfy the needs of users across specific boundaries, such as cultural, professional, organizational and geographic boundaries. Portability depends on the system of transmission, the interoperability platform which transports, translates and conveys content across a distributed environment. Portability, like usability, can be considered from either an external and contextual point of view. Portability is usually measured in terms of subjective and user-centered assessments. In the following, being portability a technological issue, we will not address it.

3.1 Quality Sub-dimensions

Several sub-dimensions can be identified in order to further refine the concept of information quality with respect to soundness and usability.

The quality of each type of content (alphanumeric structured data, text, image) can be characterized in terms of sub-dimensions that are peculiar to the considered type of content. Their names and their definition can also differ with respect to the application domain (e.g., usability of images in medical imaging differs from usability of images for e-commerce applications).

The most frequently mentioned sub-dimensions are *accuracy, consistency, completeness, accessibility, and time-related* dimensions, like *timeliness* and *currency*, since they are likely significant in every application domain. A number of these core sub-dimensions can be applied to both text-based content and images: to this aim, the definitions that have been given in the literature for a specific type of data must be adjusted to be sufficiently generic to be applicable also to the other kinds of content.

In the following, we focus on *accuracy, completeness* and *readability*, and provide general definitions according to the domain of analysis.

Defining a quality dimension implies to associate it with a set of related metrics. By the term *metrics*, we refer to the definitions given within both the ISO standard 9126-1 and the ISM3 framework [2], and hence we refer to a set of elements encompassing both a *measurement procedure*,

i.e., an algorithm that takes the element to measure and associates it with a measure (be it ordinal or interval value), and a proper *unit of measure*, i.e., the domain of values returned by the measuring procedure. In general, several metrics can be associated with each quality dimension.

Indeed, a quality dimensions can be described and characterized both at a semantic and a syntactic level. These two levels are obviously related but lead to slightly different definitions and ways to assess and measure a given dimension of a given information representation. We consider as *semantic* the level of description of a dimension that characterizes a representation with respect to a phenomenon in the represented reality of interest. We define *syntactic* the level at which a dimension characterizes a representation with respect to a reference representation (value or structure) in the same domain of representation. This twofold characterization of the concept of quality dimension is common to all types of information. For instance, an image can be assessed either intrinsically, as a stand-alone entity that can not be matched to any reference image or sound, or with respect to a reality that such content is intended to represent, e.g., the face of a person. The same can be said for a database record of a table representing the employees of an organization: its accuracy (as composition of the accuracy of each item) can be assessed both with respect to a reference vocabulary (syntactic accuracy) and with respect to the actual community of employees of that organization. The two assessments of the same entity (the record) can be quite different from each other, as in the case where, e.g., ‘John’ is an accurate item with respect to the set of all English names but no John is actually working at that particular organization.

4. QUALITY DIMENSIONS FOR ALPHANUMERIC STRUCTURED DATA

We notice that no general agreement exists either on which set of dimensions defines the quality of alphanumeric data, or on the exact meaning of each dimension. For a comprehensive discussion on this issue see [5].

Several definitions are provided for the term *accuracy*. [21] define accuracy as “the extent to which data are correct, reliable and certified”. [9] specify that data are accurate when the data values stored in the database correspond to real-world values. In [17], accuracy is defined as a measure of the proximity of a data value v to some other value v' that is considered correct. In general, two types of accuracy can be distinguished, syntactic and semantic, which we adopt in the following.

Syntactic accuracy is the closeness of a value v to the elements of the corresponding definition domain D . In syntactic accuracy, we are not interested in comparing v with its real-world value v' ; rather, we are interested in checking whether v is any one of the values in D , or how close it is to values in D . For example, $v = 'Jean'$ is considered syntactically accurate even if $v' = 'John'$.

Semantic accuracy is the closeness of value v to the corresponding true (real-world) value v' . While it is reasonable to measure syntactic accuracy using a distance function, semantic accuracy is measured with a boolean <yes, no> or a <correct, not correct> domain. Consequently, semantic accuracy coincides with the concept of *correctness*. In general, techniques assessing and improving semantic accuracy are considerably more complex than techniques addressing

syntactic accuracy.

Completeness is defined as the degree to which a given data collection includes the data describing the corresponding set of real-world objects.

In the research area of relational databases, completeness can be referred to any type of structure in the model, resulting in:

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

In all these cases, completeness is related to the meaning of *null* values defined in the model. A null value has the general meaning of *missing value*, i.e. a value that exists in the real world but is not available in a data collection. In order to characterize completeness, it is important to understand why the value is missing. A value can be missing either because it exists, but is not known, or because it does not exist, or because it is not known whether it exists (see [3]).

Let us consider the table reported in Figure 3, with attributes **Name**, **Surname**, **BirthDate**, and **Email**. If the person represented by tuple 2 has no email, 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 an 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.

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

Figure 3: Null values and data completeness

In logical models for databases, such as the relational model, there are two different assumptions on the completeness of data represented in a relation instance r . The *closed world assumption* (CWA) states that only the values actually present in a relational table r , and no other values represent facts of the real world. In the *open world assumption* (OWA) we can state neither the truth nor the falsity of facts not represented in the tuples of r .

From the four possible combinations emerging from (i) considering or not considering null values, and (ii) OWA and CWA, we consider the two most interesting cases:

1. model without null values with OWA;
2. model with null values with CWA.

In a model without null values with OWA, in order to characterize completeness we need to introduce the concept of *reference relation*. Given the relation r , the reference relation of r , called $\text{ref}(r)$, is the relation containing all the tuples that satisfy the relational schema of r , i.e., that represent objects of the real world that constitute the present true extension of the schema.

On the basis of the reference relation, the completeness of a relation r is measured in a model without null values as the fraction of tuples actually represented in the relation r , namely, its size with respect to the total number of tuples in $\text{ref}(r)$:

$$C(r) = \frac{|r|}{|\text{ref}(r)|}$$

In the model with null values with CWA, specific definitions for completeness can be provided by considering the granularity of the model elements, i.e., value, tuple, attribute and relations.

Concerning *readability*, intuitively a database, or also, its schema, is readable whenever it represents the meaning of the reality represented by the schema in a clear way for its intended use. This simple, qualitative definition is not easy to translate in a more formal way, since the evaluation expressed by the word *clearly* conveys elements of subjectivity. In models, such as the Entity Relationship model, that provide a graphical representation of the schema, called *diagram*, readability concerns both the diagram and the schema itself.

Another aspect related to readability is the property that every aspect of the real world is represented by a specific single database structure; this characteristics results in the relational model in the concept of normalization. The property of *normalization* has been deeply investigated, especially in the relational model, although it expresses a model-independent, general property of schemas. In the relational model, normalization is strictly related to the structure of functional dependencies. Several degrees of normalization have been defined in the relational model, such as first, second, third, Boyce Codd, fourth, and other normal forms. The most popular and intuitive normal form is the *Boyce Codd normal form*, *BCNF* (see [3]). A relation schema R is in BCNF if for every non trivial functional dependency $X \rightarrow Y$ defined on R , X contains a key K of R , i.e., X is a superkey of R . The interpretation of BCNF is that the relational schema represents a unique concept, with which all nontrivial functional dependencies are homogeneously associated, and whose properties are represented by all non-key attributes. For more details on the BCNF and other normal forms, see [3] and [10].

5. QUALITY DIMENSIONS FOR TEXTUAL INFORMATION

5.1 Dimensions and metrics for Textual Information

In this section, we define a set of quality dimensions that can be applied to describe relevant aspects of textual information, be it either semistructured – as in an XML file – or unstructured – as in any narrative text. For each dimension, we provide an operative definition, and for some of them we either suggest or define a quantitative metric. To this aim, we consider the main structures conveying textual informa-

tion. The smallest data structure (information unit) that we consider significant in the domain of textual information is the *word*. Words can be considered atomic elements to our practical purposes and can be either *simple* or *compound*: in the latter case, two or, less likely, more words are considered as an unity. Words are usually separated by spaces in texts; blanks can be considered as the simplest delimiters of words. The first level of words aggregation that we take into consideration is the *sentence* level. Sentences are grammatical units of one or more words, bearing syntactic relation to the words that precede or follow it. Sentences are usually separated by punctuation symbols, such as periods, semicolons and other terminal punctuation marks. We are not concerned with the meaning of sentences, nor with its either intrinsic or contextual quality; this is the object of study of the discipline called *linguistics*, whose concern is quite far from the scope of this paper. The next structure to which we apply the typical quality dimensions seen in the previous section is *text*. For the notion of text, we assume the common sense definition of set of sentences that can be considered as unitary as a single and meaningful body of matter in a manuscript, book, newspaper, etc. The next and last level we consider is that of *collection of texts*, i.e., a group of single texts that are gathered into one location, for some purpose or as a result of some process. With reference to Figure 1, we consider texts represented in any digital form as our digital information items (in which words are the atomic information units), and collections of texts as our digital information resources.

5.2 Specific dimensions and Related Metrics for Textual Information

According to [16], *accuracy* is the foundation measure of the quality of data. Moreover, its impact on the overall quality is significant, since if one makes grammatical mistakes, other aspects like conciseness and elegance are of little importance. As in the case of structured alphanumeric data – as seen in Section 4 – the concept of accuracy of an information item can be declined in terms of both a syntactic accuracy and a semantic accuracy. In the case of texts, we adapt the definition provided for alphanumeric data by first applying it to atomic information units, and then generalizing it to a whole information item (text). This means that if a word that is used in a text belongs to a reference dictionary or word list, it may be considered accurate, leaving its meaning out of consideration. Syntactic accuracy is then first assessed with respect to single words, then generalized to texts and their collections, since its conceptual assessment is scalable (e.g., by simple composition or aggregation) and its factual assessment can be carried out in quantitatively and automatic manner. Conversely, semantic accuracy regards how well a word, a text or passage, describes, i.e., is faithful to a real situation or phenomenon of the reality of interest. In this case, we should take into account the meaning of words, as well as their relationships, and also the reality they intend to represent accurately. This is an aspect that is very difficult to assess either automatically or quantitatively. That said, we propose the following definitions of accuracy related to textual information.

Syntactic Accuracy is the degree to which the reported information item is in conformance with the elements of a reference language vocabulary V . Intuitively, a quantitative metrics to assess syntactic accuracy of a textual information

item should count the number of words that are correct from the spelling point of view, and compare such number with the total number of the words contained in the text.

Semantic accuracy is the degree to which the reported information value is in conformance with the true or accepted value [6]. In this case, subjective metrics must be adopted as the only feasible approach. These metrics can differ a lot in the domain design and measure method, but they are basically based on the qualitative assessment carried out by a group of human readers, which are either expert of the subject the text is about, or not. Another approach could consider how many *specific words* are used in a sentence (or text, collection). Yet, it is not a trivial task to consider how specific a word is. Intuitively, a word is specific if, in a given thesaurus no other word specializes its intended meaning. For instance, if in a sentence it is used the term ‘thing’ to refer some object, e.g., a table, and in another describing the same situation it is used the more specific term ‘table’ or even ‘pool-table’, the latter sentence would be considered more accurate than the former. The metrics adopting this approach have to rely on a semantic lexicon for the given language used in the text, which provides for each noun, verb, adjective and adverb their synonym sets, each representing a single underlying lexical concept (e.g., such as WordNet).

Completeness refers to the extent an information item/resource has all parts or elements that are needed for a certain task or with respect to a certain schema. These two cases can be somehow related to the case of semantic and syntactic completeness, respectively. In regard to syntactic completeness, in semistructured textual resources, such as XML files, a schema exists and is explicit. In this case, the metrics to assess completeness can be borrowed from the field of alphanumeric structured data, and be based on the simple count of fields (tags) that are filled in, with respect to the total number of tags of the whole sentence, text or collection of texts. On the other hand, in case of unstructured texts, completeness can be assessed with respect to underlying requirements or explicit constraints. For instance, an abstract of a conference paper is usually something that is considered either complete or not according to a word limit fixed by the organizing committee. The convention that holds here is that the word limit is to be intended not as an upper limit, but rather as the advisable number of words needed to let the reviewers and readers understand what the paper is about. In such cases, completeness can be expressed in terms of a ratio between the number of terms used and the number of terms considered optimal for a given task (mind that in this case, also values higher than one are possible, although not always desirable).

A similar approach can be adopted also regarding the distinction between completeness under the closed world assumption and the open world assumption. In the former case, the only way to assess the completeness of a textual information item is to count the number of omissis, i.e., terms of the text deliberately left out and substituted by some place-holder (e.g., three dots, a black label). Completeness would be assessed as the ratio between the number of omissis present in a sentence (or text, collection) and the sum between this number and the overall word count. In the open world assumption, only qualitative and subjective metrics can be adopted, as in the case of semantic accuracy. In this case a panel of human experts should be consulted in order to understand whether the read text describes the object of

description in a complete manner or not.

Readability is the extent to which a text is easy to be read and understood for its targeted audience. To assess readability we may use, among others, the Gunning fog index [12], which measures the level of reading difficulty of any documents. The main idea of this method is that the higher the complexity of each sentence and the bigger the words used in it, the higher is difficulty to read the text. The resulting number is an indicator of the number of years of education a reader requires to easily understand the text at a first reading. The “standard” score is 7 or 8; and a text with score above 12 is considered hard to read for most people. The Gunning fog index may be calculated by the following algorithm:

1. Select a short passage of the text (usually around 100 words) and count the number of words. For a lengthy document, select several passages and average the Fog index.
2. Count the number of sentences.
3. Count the number of complex words, i.e. words with three or more syllables, not including proper nouns (for example, ‘Frederick’), compound words like “newspaper”, or common suffixes such as -es, -ed, or -ing as a syllable, or familiar jargon.
4. Calculate the average sentence length (i.e., divide the number of words by the number of sentences).
5. Calculate the percentage of complex words.
6. Add the average sentence length and the percentage of complex words, and multiply the result by 0.4

The complete formula is as follows:

$$0.4 * \left(\left(\frac{\text{words}}{\text{sentence}} \right) + 100 \left(\frac{\text{complex words}}{\text{words}} \right) \right) \quad (1)$$

6. QUALITY DIMENSIONS FOR IMAGES

6.1 Dimensions and metrics for images

In general, image quality dimensions can be assessed either for an image seen in isolation (e.g. readability) or for an image seen together with the reference (e.g. semantic accuracy). Depending on the image data and application, quality dimension assessment can be done by psychological experiments involving human observers or computing suitable metrics directly from the digital image, these metrics can be eventually combined as proposed in [4] and [11]. Standard psychophysical scaling tools for measuring subjective image quality are available and described in specific standards, such as ITU-R BT.500-11 [20], [13]. The involvement of real people that view the images to assess their quality requires that all the factors that influence perception are taken into account and strict protocols are adopted. Subjective image quality assessment necessarily involves taking into account both the Human Vision System characteristics and the image rendering procedure, subjects’ characteristics and the perceptual task. Subjective image quality may be assessed only when the image is viewed by an observer. Objective image quality measures not requiring human interaction can be broadly classified in two classes: signal-based



Figure 4: Low syntactic accuracy.

metrics, such as the root mean square error; and perceptually based metrics, which relate image signals to perceived quality. These metrics include simplified models of the human vision system, such as the visible differences predictor model [8].

6.2 Specific Dimensions and Related Metrics for Images

Following [5], we can define the *syntactic accuracy* of an image as its closeness to an image in the chosen application domain. For example, if the application is a biometric authentication system based on face detection and recognition, any image that does not contain a detectable/recognizable face should be considered outside the application domain and thus discarded. Syntactic accuracy can be assessed either by visual inspection or by pattern recognition techniques [14]. An example is shown in Figure 4, related to a face-based biometric recognition system. In such systems the detection of a face in an image can be used to discriminate between syntactically accurate images (first three images from the left) and other images (right). In this case, automatic determination of the syntactic accuracy of the images can be accomplished using a face detection algorithm [22]. In the rightmost image the face recognition fails.

Image semantic accuracy can be defined as the degree of matching (fidelity) of the digital image with respect to the corresponding (measured) external reference in the reality of interest, i.e. the original scene or source data we want to represent. If the image is synthesized from non-image data, its semantic accuracy is obviously related to the semantic accuracy of the data itself. Figure 5 shows an example of images with different levels of semantic accuracy (fidelity) with respect to the reference image. The background of the middle image is noisy while the color and the logo text of the left image are wrong. In this figure, the middle and right images are examples of low semantic accuracy with respect to the reference image (depicted on the left).

In some cases, semantic accuracy could be assessed by a human viewer without requiring the availability of the external reference. In these cases we can define a "reduced" semantic accuracy as the degree of apparent match of the image with the viewer's internal references [15]. Examples of image requiring a high degree of naturalness are those seen on journals. Naturalness plays a fundamental role when the image to be evaluated does not exist in the reality, such as in virtual reality domains. In Figure 6 we see two examples of images, that, despite being faithful with respect to the original scene (left) or the source data (right), lack of naturalness.

In regards to completeness, we can adapt the definition of completeness provided for alphanumeric data and texts as follows: a digital image is *complete* if it depicts all the



Figure 5: Low semantic accuracy.



Figure 6: Lack of naturalness.

information that it must convey. There are several causes for the lack of completeness and they can be related to i) acquisition or production process that generate the image (e.g. Figure 7) or to ii) the represented phenomenon (real or synthesized) being intrinsically incomplete (e.g. Figure 8).

Figure 7 shows a subject acquired with a 3D laser scanner. It can be seen that the image exhibits several holes where the scanner has been not able to correctly acquire the model surface. The 3D representation of the model is thus incomplete.

Figure 8 shows a partially occluded face. The incompleteness of this image is related to the incompleteness of the information in the acquired scene. In this case the image depicts the entire scene's visible information but the scarf

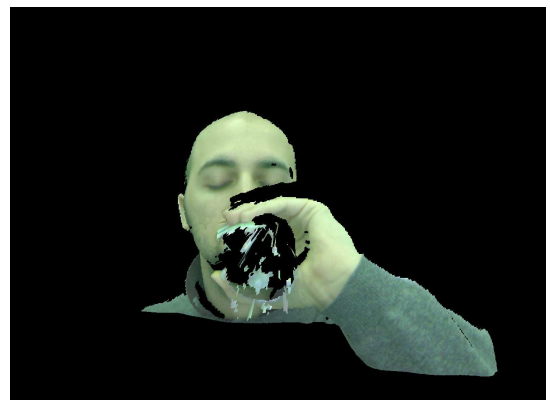


Figure 7: First example of incompleteness.



Figure 8: Second example of incompleteness.

occludes other information concerning the traits of the subject's face.

The definition of completeness given above can be applied both to a single image as well as to a set of images intended to capture a single phenomenon. For example, the acquisition of a large painting at high resolution is usually done by topologically subdividing it into small tiles that are individually acquired at the required resolution. An example of this acquisition technique can be found in [1] where the mural painting "The Last Supper" of Leonardo da Vinci was acquired using 1,677 tiles resulting in a virtual image of size 172,181 x 93,611 pixels. This set of tiles must be complete to be properly combined to provide the whole painting.

We finally define *readability* of an image as the lack of distortions or artifacts that reduce the accessibility of its information contents. Some of the most frequent artifacts are: *blurriness*, *graininess*, *blockiness*, *lack of contrast* and *lack of saturation*. To detail each of these terms falls out of the scope of the paper. In Figure 9, we provide some examples of artifacts that may reduce the readability of an image. The readability can be assessed either directly by human visual inspection or indirectly by automatically estimating the presence and the strengths of these artifacts. The estimation's process requires the modeling of the visual effects that an artifact produces on the image in terms of distortions of low level visual features, such as color, edge, texture.

7. CONCLUSIONS AND FUTURE WORK

In this paper we made a first attempt to establish a unified approach to information quality for heterogeneous types of information. We have focused on three quality dimensions, namely, accuracy, completeness and readability. Much work has to be done to formally express such model in terms of a complete classification of quality dimensions that encompass all types of qualities mentioned in the literature [5], and to extend the approach to other relevant information types such as maps [18] and sounds [7].

8. REFERENCES

- [1] Available on line: www.haltadefinizione.com/en/cenacolo/index.asp.
- [2] V. Aceituno. ISM: Information security management maturity model - handbook. Technical report, ISM3 Consortium, 2007.
- [3] P. Atzeni and V. de Antonellis. *Relational Database Theory*. The Benjamin /Cummings Publishing Company, Inc., 1993.



Figure 9: Low readability.

- [4] C. J. Bartleson. The combined influence of sharpness and graininess on the quality of colour prints. *Journal Photog. Science*, 1982.
- [5] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies, Techniques*. Springer Verlag, 2006.
- [6] D. Becker, W. McMullen, and K. Hetherington. A flexible and generic data quality metamodel. *International Conference on Information Quality*, 2007.
- [7] J. Blauert and U. Jekosch. *Sound-Quality Evaluation - A Multi-Layered Problem - EAA-Tutorium on Aurally Adequate Sound-Quality Evaluation*. EAA, Antwerp, Netherlands, March 1996.
- [8] S. Daly. The visible difference predictor: an algorithm for the assessment of image fidelity. *Digital images and Human vision*, pages 179–206, 1992.
- [9] P. H. D.P. Ballou. Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2), 1985.
- [10] R. Elmasri and S. Navathe. *Foundamentals of database systems, Fifth Edition*. Addison-Wesley Publishing Company, 1994.
- [11] P. Engeldrum. Psychometric Scaling: Avoiding the Pitfalls and Hazards. pages 101–107, 2001.
- [12] R. Gunning. *The Technique of Clear Writing*. McGraw Hill, New York, NY, 1952.
- [13] ITU. Methodology for the Subjective Assessment of the Quality for Television Pictures. ITU-R Rec. BT. 500-11, 2002.
- [14] R. Jain, S. Antani, and R. Kasturi. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4):945–965, 2002.
- [15] T. J. W. M. Janssen and F. J. J. Blommaert. Predicting the usefulness and naturalness of color reproductions. *Journal of Imaging Science and*

- Technology*, 44:93–104, 2000.
- [16] J. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.
 - [17] T. Redman. *Data Quality for the Information Age*. Artech House, 1996.
 - [18] Shekhar, Shashi, Xiong, and Hui, editors. *Encyclopedia of GIS*. Springer Verlag, 2008.
 - [19] D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Communications of ACM*, 40(5):103–110, 1997.
 - [20] L. Thurstone. A law of comparative judgement. *Psychological Review*, 34:273–286, 1927.
 - [21] R. Wang and D. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 1996.
 - [22] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.