Data Quality under the Computer Science perspective

Monica Scannapieco
Universita' di Roma, "la Sapienza", Rome, Italy
IASI-CNR, Rome, Italy
monscan@dis.uniroma1.it

Tiziana Catarci
Universita' di Roma, "la Sapienza", Rome, Italy
catarci@dis.uniroma1.it

Abstract: *La qualità dei dati è una tematica affrontata in ambito statistico, gestionale, informatico, insieme a molti altri settori scientifici. Il presente articolo considera il problema della definizione della qualità dei dati dal punto di vista informatico. Sono comparate alcune proposte di dimensioni (o caratteristiche) che contribuiscono alla definizione della qualità dei dati e viene introdotta una definizione "base" di tale concetto. E' inoltre illustrata una classificazione che ha l'obiettivo di guidare nella scelta della definizione di qualità dei dati maggiormente adeguata alle proprie esigenze.*

1 INTRODUCTION

The term "data quality" is used with reference to a set of characteristics that data should own, such as accuracy, i.e. a degree of correctness, or currency, i.e. a degree of updating. A good quality level is very important in order to provide services in both public and private contexts. Let us consider as an example a recent Italian law that allows Italian people resident outside Italy to vote; in order to make this law effective, it is necessary to improve the currency of the addresses of such citizens, because, according to a recent assessment, most of them are outdated [5].

Data quality has been addressed in different research areas, mainly including statistics, management and computer science. The statistics researchers were the first to investigate some of the problems related to data quality by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 60's [3]. The management research began at the beginning of the 80's; the focus was on how to control data manufacturing systems in order to detect and eliminate data quality problems (see as an example [1]). Only at the beginning of the 90's, computer science researchers began considering the data quality problem, specifically how to define, measure and improve the quality of electronic data, stored in databases, data warehouses and legacy systems. Many results concerning data quality have been achieved in

computer science since those years. In almost all emerging systems and technologies, data quality has an important role. In single information systems, methodologies to model specific quality requirements in relational databases [15] and to improve process quality [13] have been proposed. When multiple data sources have to be integrated, the need for a good data quality far increases, such as in data warehouses [6] or in Cooperative Information Systems (CIS) [7]. An example of a CIS is provided by an e-Government scenario in which public administrations cooperate in order to fulfill service requests from citizens and enterprises; administrations very often prefer asking citizens for data, rather than other administrations that have stored the same data, because their quality is not known.
This paper focuses specifically on the definition of data quality as it has been considered in the computer science field. The aim of the paper is to give the reader a background on the different definitions for data quality that have been proposed since the 90's; some correspondences among the different definitions are also outlined and a classification of the different proposals for a data quality definition is suggested. The classification can be seen as a guide when choosing the most appropriate quality definitions for one's own requirements.

2 DATA QUALITY AS A SET OF DIMENSIONS

Data quality has been defined as “fitness for use” [17], with a specific emphasis on its subjective nature. Another definition for data quality is “the distance between the data views presented by an information system and the same data in the real world” ([9], [14]); such a definition can be seen as an “operational definition”, although evaluating data quality on the basis of comparison with the real world is a very difficult task.
Generally speaking, data quality can be related to a set of “dimensions” that are usually defined as quality properties or characteristics. Examples of dimensions are accuracy, completeness and consistency. An intuitive understanding of such dimensions is suggested by the following examples; they refer to a record *Citizen*, with fields *Name*, *Sex* and *Email*, shown in Figure 1.

CITIZEN		
Name	**Sex**	**Email**

Figure 1: Example record.

If *Name* has a value *Mke*, while Mike is the correct value according to a dictionary of English names, this is a case of low accuracy.
An example of low completeness is provided by considering *Email*; a null value for *Email* may have different meanings, that is *(i)* the specific citizen has no e-mail address, and therefore the field is inapplicable (this case has no impact on completeness), or *(ii)* the specific citizen has an e-mail address which has not been stored (in this case the degree of completeness is low).
As an example of consistency, let us consider the values of the fields *Name* and *Sex*. If *Name* has a value that is *John* and the value of *Sex* is *Female*, this may be a case of low consistency.

It is worth noting that:

- In the literature, there is no agreement on the set of the dimensions characterizing data

quality. Many proposals have been made, but no one has emerged above the others and has established itself as a standard.

- Even if some dimensions are universally considered as important, there is no agreement on their meanings. In different proposals, the same name is often used to indicate semantically different things (as well as different names are needed for the same thing). In the remaining of this section, we first list the main proposals for a set of dimensions defining data quality (Section 2.1), and then we compare such proposals (Section 2.2).

2.1 The proposals for data quality dimensions

In 1995, a survey of the proposed sets of dimensions characterizing data quality was published [16]. Since then, many other proposals have been developed; among them we have selected six ones that are representative of different contexts. We introduce them in the following:

- **WandWang96** [14]. It is based on a formal model for information systems. Data quality dimensions are defined by considering mapping functions from the real world to an information system. As an example, inaccuracy of data means that the information system represents a real world state different from the one that should have been represented. As another example, the completeness dimension is defined as a missing mapping from real world states to the information system states. A total of 5 dimensions are proposed: accuracy, completeness, consistency, timeliness, and reliability.
- **WangStrong96** [17]. The proposal derives from an empirical study. Data quality dimensions have been selected by interviewing data consumers. Starting from 179 data quality dimensions, the authors selected 15 different dimensions.
- **Redman96** [12]. The proposal groups data quality dimensions into three categories, corresponding to the conceptual view of data, the data values and the data format respectively. 5 dimensions are proposed for the conceptual view, 4 dimensions for the data values and 8 dimensions for the data format.
- **Jarke99** [10]. The proposal was made in the context of the European Research Project DWQ, Foundations of Data Warehouse Quality. The overall project objective is to guide data warehouse design activities. In this context, specific data quality dimensions are proposed. The dimensions are classified according to the roles of users in a data warehouse environment, namely: 6 dimension for Design and Administration Quality, 6 dimensions for Software Implementation Quality, 5 dimensions for Data Usage Quality and 5 dimensions for Data Stored Quality.
- **Bovee01** [2]. Following the concept of data quality as “fitness for use”, the proposal includes 4 dimensions (with some sub-dimensions). Data “fit for use” whenever a user: 1) is able to get information (Accessibility); 2) is able to understand it (Interpretability); 3) finds it applicable to a specific domain and purpose of interest (Relevance); 4) believes it to be credible (Credibility).
- **Naumann02** [8]. The proposal defines quality dimensions specific for integrated Web Information Systems [4]. It considers 4 categories for a total of 21 dimensions. The four categories are: content-related, concerning the actual data that are retrieved; technical, concerning aspects related to the source, the network and the user; intellectual, related to subjective aspects of the data source; instantiation-related, concerning the presentation of data.

2.2 Comparison of data quality dimension proposals

In comparing the different proposals for data quality dimension sets, we highlight, in Figure 2 and Figure 3, two types of correlations among them, namely:

- In Figure 2, we show how different proposals use the same name for dimensions with **Different (D)** or **Similar (D%)** meanings. The letter **S** is used to indicate same names and same meanings for a dimension in the different proposals; this is outlined in order to consider which proposals include the same dimensions.
- In Figure 3, we see how different names for dimensions with **Similar (D%) or Same (S)** meanings are used.

According to Figure 2, accuracy and completeness are the only dimensions defined by all proposals. Besides these two specific dimensions, consistency-related dimensions and time-related dimensions are also taken into account by all proposals. Specifically, consistency is typically considered at instance level (consistency dimension) or at format level (representational consistency). Time-related quality features are mainly caught by the timeliness dimension. Also interpretability is considered by most of the proposals, both at format and schema level.
Each of the remaining dimensions is included only by a minority of proposals. In some cases there is a complete disagreement on a specific dimension definition, such as for reliability.
For detailed comments on each raw of the table shown in Figure 2, the reader can refer to the Appendix.
In Figure 3, we show similarity between dimensions that are named differently in the various proposals:

- In Figure 3.a, clarity of definition as defined in **Redman96** is similar (D%) to interpretability as defined in **WangStrong96**, **Bovee01** and **Naumann02**.
- In Figure 3.b, accessibility as defined in **WangStrong96**, **Jarke99** and **Bovee01**, i.e. how much data are available or quickly retrievable, is the same (S) as obtainability of values in **Redman96**.
- In Figure 3.c, correctness as defined in **Jarke99**, i.e. proper comprehension of the entities of the real world, is the same (S) as comprehensiveness in **Redman96.**
- In Figure 3.d, minimality as defined in **Jarke99**, i.e. the degree up to which undesidered redundancy is avoided, is the same (S) as minimum redundancy as defined in **Redman96**.

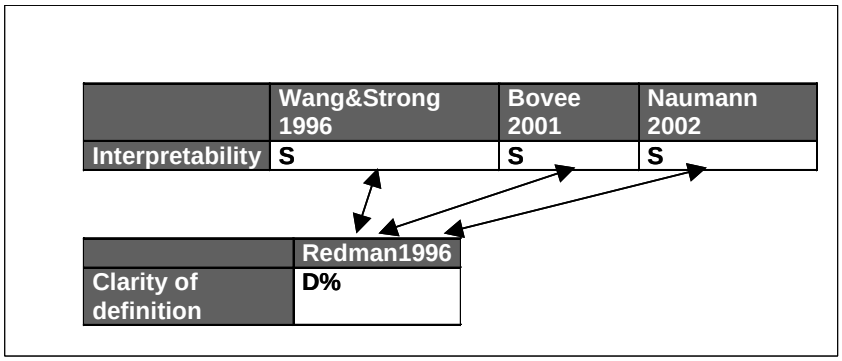
On the basis of the correlations in Figure 2 and Figure 3, it is possible to sketch a basic definition for data quality, i.e. a definition that includes features considered by the majority of the proposals.
We define *data quality as a set of dimensions including accuracy, completeness, consistency (at format level and at instance level), timeliness, interpretability and accessibility.*
In fact, Figure 2 shows that accuracy, completeness, consistency, timeliness and interpretability are dimensions shared by most of the proposals. Combining Figure 2 and Figure 3, also accessibility has to be included in the basic set.
The reader should notice that for each proposal only the dimensions that are shared by at least another proposal have been considered. As an example, the list of **Jarke99**'s dimensions includes six further dimensions besides the ones shown in Figure 2, that are not shared by any of the other

proposals.

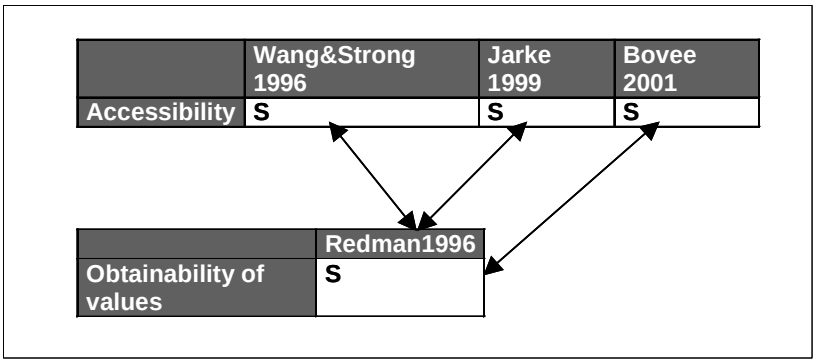
	WandWang 1996	WangStrong 1996	Redman 1996	Jarke 1999	Bovee 2001	Naumann 2002
Accuracy	S	S	S	S	S	S
Completeness	S	D	S	D%	S	S
Consistency	S		D%	S	S	
Representational Consistency		S	S			S
Timeliness	S	S		S	S	S
Currency	S		S	S	S	
Volatility	S			S	S	
Interpretability		S	D%	D%	S	S
Ease of Understanding/ Understandability		S				S
Reliability	D			D		
Credibility				D	D	
Believability		S				S
Reputation		S				S
Objectivity		S				S
Relevancy/ Relevance		S	S		D%	S
Accessibility		S		S	S	
Security/ Access Security		S		S		S
Value-added		S				S
Concise representation		S				S
Appropriate amount of data/amount of data		D	D			D
Availability				S		S
Portability			D	D		
Responsiveness/ Response Time				S		S

Figure 2: Correspondences among dimensions with same name.

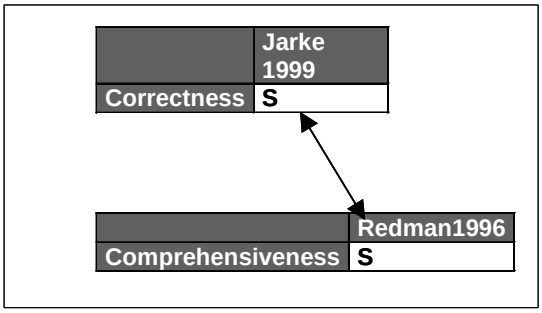
In summary, though there are several different dimensions in the various proposals, it is possible to single out a few of them that basically define the concept of data quality. All other dimensions included in the proposals either capture secondary features or are more context-dependent (i.e. very specific). This latter case will be deeply discussed in the next section.



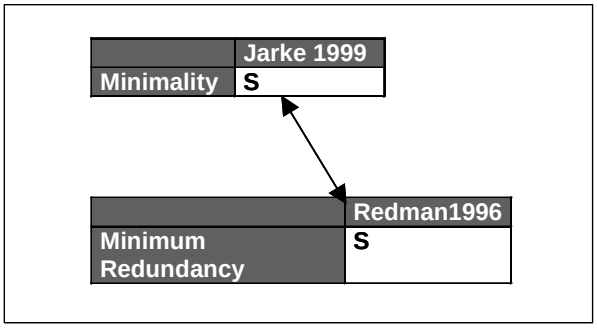
(a)



(b)



(c)



(d)

Figure 3: Correspondences among dimensions with different names.

3 A CLASSIFICATION OF THE DATA QUALITY DIMENSIONS PROPOSALS: WHEN TO USE WHICH PROPOSAL

Though having provided for a basic definition of data quality in Section 2, there are many cases in which more specific sets of dimensions are needed. The aim of this section is to classify the proposals for data quality dimensions in order to have some methodological suggestions guiding the choice of the best proposal according to one's own requirements.

We classify the proposals described in Section 2 according to four features:

- *Approach to dimensions definition*, i.e., what is the process followed to define the set of dimensions. We consider three types of possible approaches to dimension definition: **Theoretical**, **Empirical** and **Intuitive**. A theoretical approach means that a formal model or theory is proposed in order to define or justify the proposed dimensions. An empirical approach means that the set of dimensions is constructed starting from experiments, or interviews and questionnaires. In the intuitive approach, dimensions are simply defined according to common sense.
- *Modeling view on data*, i.e., which data perspective has been taken in order to define quality dimensions. The considered perspectives are three: **Schema**, **Instance** and **Format**. Schema refers to the intensional view of data, such as the definition of a table in the relational model; the instance view is related to the extension of data, i.e. to actual values; the format view is related to how data are represented.
- *Measurement view on data*, i.e., how data are analyzed in order to evaluate their quality. Three different measurement views can be considered: **Process**, **System** and **Product**. The process view considers how the processes that produce data affect the data quality. As an example, it is possible to consider the way in which a data entry process is carried on and how it affects the quality of the stored data. The system view is related to the consideration of the whole information system as influencing the quality of data. As an example, if a distributed information system is considered, some specific attention should be paid to time related dimensions and to how their values are affected by data exchanges among different nodes. Instead the product view is specifically related to data and to the user perception of their quality. More specifically, this view considers information as a product to be delivered to consumers and typically includes subjective dimensions like interpretability or understandability, which need to be evaluated by the final consumer of the information.
- *Context dependence*, i.e., if the proposal is tailored to a specific application context or is general purpose. The possible values of this classification variable are simply **Yes** or **Not**.

In Figures 4.a and 4.b, the positions of the various proposals according to the described variables are shown. Figure 4.a shows the positions of each proposal in the plan *Approach to Dimension Definition-Modeling View on Data*. Notice that the only proposal covering all modeling views on data is **Redman96**. **Jarke99** is specific for the data warehouse context, and **Naumann02**, for web integration systems. Figure 4.b shows the position of each proposal in the plan *Approach to Dimension Definition-Measurement View on Data*. Notice that no proposal covers all measurement views on data.

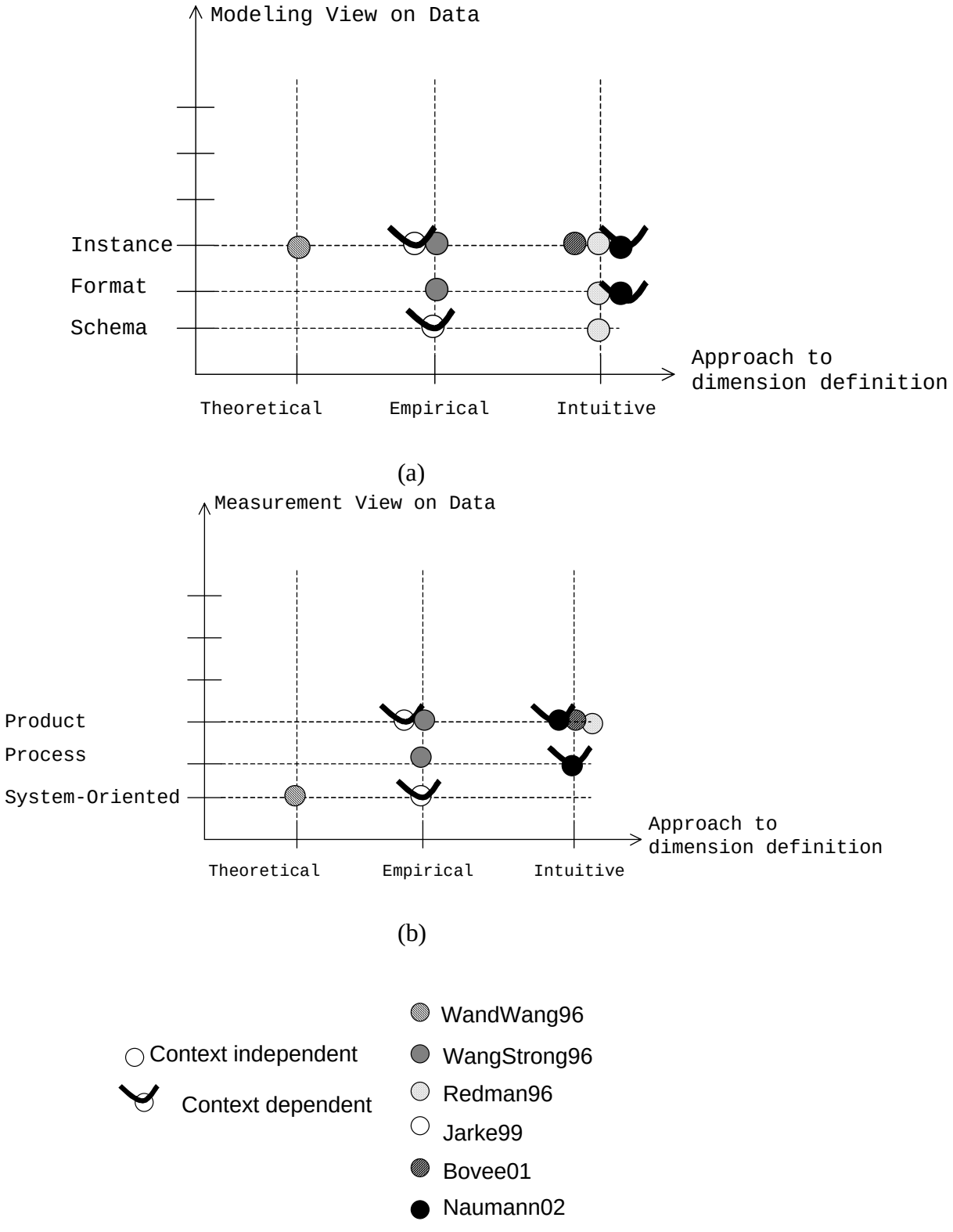


Figure 4: A classification of dimension proposals.

The various *approaches to dimension definition* give useful information for a good choice of the best set of quality dimensions. An intuitive proposal may be sufficient if a general introduction to the data quality problem is required. Whereas, when data quality information has to guide strategic decisions, an empirical or theoretical approach may be more adequate.
The different *modeling views on data* also help in focusing on the right abstraction level to be adopted. As an example, if unstructured information is considered, such as free text documents, no proposal concentrating on schemas may be useful. As another example, when considering images, a proposal specifically concerning format may be most appropriate.
With reference to *measurement views*, the process view allows for focusing on quality improvement inside an organization, by detecting the key causes of poor quality. Organizational decisions may be taken, for example concerning the enactment of a Business Process Reengineering (BPR) activity. The system view becomes very important when a particular information system supports a given enterprise or organization. The product view is especially important when data are the real product of an enterprise, such as in public administrations that have as a main task to manage data about citizens and enterprises.
Finally, the *context dependence* is of great importance. Indeed, if a specific proposal fits the one's own context, it is undoubtedly the right one to choose.
As an example of how to combine the different variables in order to choose a specific proposal of dimensions, let us consider the Italian e-Government scenario. Current Italian e-Government projects aim at creating a cooperative information system among public administrations. Autonomous administrations must be able to cooperate with other administrations since they do not have complete control over data and services needed to reach their own goals. In such a scenario, if an administration has to request data from another, it would like to be guaranteed about the quality of the data provided by this latter. An assessment of the quality provided by each administration is thus necessary in order to enable data exchanges among cooperating organizations. Let us consider some examples of quality requirements that drive the choice of quality dimensions:

- Specific focus on data values, as they are the ones actually exchanged among the different administrations.
- Focus on data as a product to be delivered by the nationwide CIS to citizens and enterprises.
- Rigorous process to be followed in dimension definition.

According to such requirements a good set of dimensions can be provided by **WangStrong96**. In fact, this proposal focuses on instance view of data, thus addressing the requirement related to data values, and on a measurement view of *product* type. The empirical approach to dimension definition is also adopted thus addressing the need of a rigorous process in dimension definition. The fact that **WangStrong96** also includes a measurement view of data of *process* type (see Figure 4.b) makes this proposal suitable even for an improvement activity that could be engaged by single administration on their own data. As an example, an administration may find out that some typographical errors in the data it disseminates are due to the lack of an automatic data entry process. The solution may be a reengineering of the data entry process in which data are automatically loaded, for example from XML files sent by a different administration.
The proposed classification aims at guiding the designer in choosing data quality dimensions that fits his/her application needs. It is possible (and it is also very probable) that no proposal matches exactly one’s requirements. In such cases, the role of the classification is to help in pointing out

the specific deficiencies of each proposal.
Let us consider again our example of the Italian CIS among public administrations. CIS's are information systems with a lot of specific characteristics; this suggests the usefulness of a set of dimensions also taking into account a *system* perspective. Let us suppose that the Department of Finance asks a City Council for some data related to a citizen's family composition, in order to enact a tax assessment. Such data may arrive late and may cause the tax assessment process to be postponed because of a lack of system availability, a need of secure transmission, delays of the networks etc. Therefore, besides the dimensions strictly related to the quality of data, some other dimensions might need to be considered as directly impacting the quality of the data as perceived by the final consumer. In our example, the timeliness of the data arrived at the Department of Finance is affected by system quality problems. **WangStrong96** does not allow for considering a measurement view on data of system type, so the choice of this proposal for the Italian CIS should be integrated with some specific dimensions explicitly taking into account system characteristics.

4 REFERENCES

[1] Ballou D.P., Pazer H.L.: Modeling data and process quality in multi-input, multi-output information systems, *Management Science*, vol.31, no.2, 1985.

[2] Bovee M., Srivastava R. P., Mak B.R.: A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. In *Proceedings of the 6th International Conference on Information Quality*, Boston, MA, 2001.

[3] Fellegi I.P., Sunter A.B.: A Theory for Record Linkage. *Journal of the American Statistical Association*, vol.64, 1969.

[4] Isakowitz T., Bieber M., Vitali F. (eds): Special issue on Web Information Systems, *Communications of the ACM*, vol.41, no.7, 1998.

[5] Italian National Newspaper: "Il Sole 24 Ore". No. 115, published on April 29th 2002.

[6] Jarke M., Lenzerini, Vassiliou Y., Vassiliadis P.*: Fundamentals of Data Warehouses*. Springer Verlag, 1999.

[7] Mecella M., Scannapieco M., Virgillito A., Baldoni R., Catarci T., Batini C.: Managing Data Quality in Cooperative Information Systems. In Proceedings of the Tenth International Conference on Cooperative Information Systems (CoopIS 02), Irvine, CA, 2002.

[8] Naumann F.: *Quality-Driven Query Answering for Integrated Information Systems*, LNCS 2261, 2002.

[9] Orr K.: Data Quality and Systems Theory. In *Communications of the ACM*, vol. 4, no. 2, 1998.

[10] Quix C., Jarke M., Jeusfeld M.A., Vassiliadis P.: Architecture and Quality in Data Warehouses: an extended Repository Approach, *Informayion Systems*, vol.24, no.3, 1999.

[11] Rahm E., Do H. H.: Data Cleaning: problems and current approaches. IEEE Data Engineering bulletin, Vol. 23, no.4, 2000.

[12] Redman T.C.: *Data Quality for the Information Age*. Artech House, 1996.

[13] Shankaranarayan G., Wang R. Y. and Ziad M.: "Modeling the Manufacture of an Information Product with IP-MAP". In Proceedings of the *6th International Conference on Information Quality*, Boston, MA, 2000.

[14] Wand Y., Wang R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations. *Communication of the ACM*, vol. 39, no. 11, 1996.

[15] Wang R.Y., Kon H.B., Madnick S.E.: Data Quality Requirements: Analysis and Modeling. In *Proceedings of the 9th International Conference on Data Engineering (ICDE '93*),Vienna, Austria, 1993.

[16] Wang R.Y., Storey V.C., Firth C.P.: A Framework for Analysis of Data Quality Research. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 7, No. 4, 1995.

[17] Wang R.Y., Strong D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of *Management Information Systems*, vol. 12, no. 4, 1996.

5 APPENDIX

This section details the relations among the dimension definition and meaning in the various proposals shown in Figure 2.

Accuracy is defined in all proposals as the degree of correctness of a value when comparing with a reference one.

Completeness is defined as the degree of presence of data in a given collection in almost all proposals. In **WangStrong96,** a different user-dependent definition is provided, i.e. the extent to which data are of sufficient breadth, depth and scope for the task at hand. In **Jarke99,** a similar definition is provided, but completeness is defined also at schema level, rather than at instance level only.

Consistency is defined by all the proposals but **WangStrong96** and **Naumann02**, which instead define representational consistency. In all definitions it is the consistency among different data values (e.g. Sex and Name). Only in **Redman96**, consistency is defined at conceptual level, format level and at instance level.

Representational consistency has the same meaning in all the proposals that define it, i.e. the extent to which data are always presented in the same format.

Timeliness, *currency* and *volatility* are all time-related dimensions. Timeliness is defined as the extent to which data are timely for their use. Timeliness can be defined in terms of currency (how recent are data) and volatility (how long data remains valid). In **Redman96** only currency in defined, while in **WangStrong96** and **Naumann02** only timeliness.

Interpretability is related to the format in which data are specified, including language spoken, units, etc. and to the clarity (non-ambiguity) of data definitions. This dimension is very similar to *Understandability/Ease of understanding*, though both **WangStrong96** and **Naumann02** propose Interpretability and Understandability/Ease of understanding as distinct dimensions. In **Redman96** two distinct dimensions are proposed: interpretability and clarity of definitions, a subdimension of content. Also in **Jarke99**, schema and data interpretability are proposed.

Reliability, *Credibility*, *Believability*, *Reputation* and *Objectivity* also are very related dimensions.

In **WandWang96**, reliability indicates whether the data can be counted up to convey the right information. A completely different meaning is provided in **Jarke99**, in which it is the frequency of failures of a system, its fault tolerance.

In **Jarke99,** credibility is the credibility of the source that provided information. A different meaning is given to credibility in **Bovee01**, i.e. how much information is accurate, complete, consistent and non-fictiousness.

Believability, Reputation and Objectivity are defined both in **WangStrong96** and in **Naumann02**. Believability is the extent to which data are accepted or regarded as true, real and credible; reputation is the extent to which data are trusted or highly regarded in terms of their

source or content. Objectivity is the extent to which data are unbiased (unprejudiced) and impartial.
Relevancy or *Relevance* is a dimension included in all the proposals but **WandWang96**. Relevancy considers how data are relevant for the task at hand. Only in **Bovee01** a bit different dimension is provided, i.e. data are relevant if are timely and satisfy user-specified criteria, thus explicitly including a time related concept.
Accessibility expresses how much data are available or quickly retrievable. There's an agreement on its definition as well as on the definition of Access Security or Security as implying access to data to be kept secure.
Value-added and *concise representation* are two dimensions defined both in **WangStrong96** and **Naumann02**. Value-added is related to how much data provide benefits for the users. Concise representation is related to how much data are compactly represented.
In **WangStrong96**, *appropriate amount of data* is the extent to which the quantity or the volume of data is appropriate. A more specific definition is provided by **Naumann02**, i.e. the size of the query result measured in bytes. In **Redman96,** the same name is used to refer to the appropriateness of the format.
Availability is only defined in **Jarke99** and in **Naumann02** as the availability of a data source or a system.
Portability is defined as system independence of the data format **Redman96**; instead in **Jarke99** it has a quite different meaning, i.e. it represents system independence of the software managing data.
Responsiveness or *Response time* is defined as the time interval between the submission of a query and the answer, in both **Jarke99** and **Naumann02**.