

A FRAMEWORK AND A METHODOLOGY FOR DATA QUALITY ASSESSMENT AND MONITORING

(Practice-Oriented)

Carlo Batini

Università di Milano Bicocca, Italy
batini@disco.unimib.it

Daniele Barone

Università di Milano Bicocca, Italy
barone@disco.unimib.it

Michele Mastrella

FutureSpace, Italy
michele.mastrella@futurespace.it

Andrea Maurino

Università di Milano Bicocca, Italy
maruino@disco.unimib.it

Claudio Ruffini

Augeos srl
claudio.ruffini@augeos.it

Abstract: Data quality (DQ) is emerging as a new relevant area for the improvement of the effectiveness of organizations. Despite the consequences of poor quality of data are often experienced in everyday life of enterprises, very few organizations adopt specific methodologies for assessing and monitoring quality of their data. In this paper we present the first results of an Italian project whose goal is to produce an enhanced version of well known approaches to Basel II operational risk evaluation, with a significant relevance to information and data quality, and its effects on operational risk. In particular in this paper we focus on the definition of an assessment methodology and a supporting tool for DQ.

Key Words: Data quality, Data assessment, Methodology, Tool for data assessment

INTRODUCTION

The consequences of poor quality of data are often experienced in everyday life of enterprises, but, often, without making the necessary connections to their causes. Data quality has serious consequences, of far-reaching significance, for the efficiency and effectiveness of businesses. The report on data quality of the Data Warehousing Institute [10] estimates that data quality problems cost U.S. businesses more than 600 billion dollars a year.

In the private sector, application providers and system integrators are experiencing the role of DQ in their own products and services. IBM's recent (2005) acquisition of Ascential Software, a leading provider of data integration tools, highlights the critical role data and information stewardship plays in the enterprise. In Italian, Orme means tracks, footsteps, marks. The ORME project has been granted by the Italian Ministry of Economic Development, and involves several businesses and research groups in the

production of a methodology and a framework that add value to existing methodologies and tools proposed for Basel II operational risk evaluation [3]. Businesses involved in ORME (Futurespace, PEG Group, and Augeos), in a joint effort with the SeQuOIaS (Service and Quality Oriented Information Systems) research group of University of Milan Bicocca are conceiving an enhanced version of well known approaches to Basel II operational risk evaluation. The novelty is focused in the relevance assigned to information and data quality, and its effects on operational risk. In this paper, we present the first result of this activity: ORME-DQ a methodology and an associated framework for the assessment of data quality in an organization.

The paper is organized as following: in the next Section the phases and steps of the assessment methodology are provided. In Architecture Section we describe the architectural framework supporting the methodology, and we describe, by means of ER schema, the main information that the ORME-DQ methodology manages. Finally in Related Section most important assessment methodologies are compared and then we draw our conclusion and future work.

ORME-DQ: an Assessment methodology

Measuring data quality in a single organization is a complex task. The methodology we present is derived by the one presented in [5] and it is a reasonable balance between completeness and the practical feasibility of the data quality process. Figure 1 provides a description at a glance of ORME-DQ.

Phase 1: DQ Risk Prioritization

1. Reconstruct the state and meaning of most relevant databases, data flows, business processes, loss events related to Organizational Units, and build correlation *matrixes*

Phase 2: DQ Risk Identification

1. Loss event profiling and evaluation of economic losses
2. Selection of critical processes
3. Selection of critical data sets & data flows

Phase 3: DQ Risk Measurement

1. Qualitative and Quantitative Assessment of data quality
2. Approximate evaluation of loss events conditional probability, given the historical series of DQ dimensions quantitative assessments

Phase 4: DQ Risk Monitoring

1. Evaluate the DQ dimension values periodically and send alerts when they are less than predefined target values

Fig. 1. Phases and steps of methodology.

The innovative aspects of ORME-DQ are:

- The use of a wide number of knowledge's sources for measurement decisions, ranging from

organizations, to business processes, data sets, exchanged data flows, etc.

- Risk measurement is performed through a mix of qualitative and quantitative assessment.
- The quantitative assessment is focused on the most relevant quality dimensions, the 4DQ dimensions, namely accuracy, completeness, currency and consistency [5].
- Economic losses are estimated adopting a detailed classification of costs related to low data quality shown in Figure 2 and also by adopting an approximate technique, the discriminant analysis (see below), based on the construction of discrete category subsets of predictor values in a multidimensional predictor space. The classification of costs in Figure 2 is much richer than usual checklists adopted in information management techniques and it is the result of the integration of three very detailed costs classifications [6,15,7].

Phase 1: DQ Risk Prioritization

In phase 1, all the relationships between organization units, processes, services, and data, if not known, are reconstructed. Similarly to what happens in information system planning methodologies, at the beginning of the DQ process we reconstruct a model of the most relevant relationships between organizations or organizational units and data used and exchanged. The goal of this phase is to provide a picture of the main uses of data, of providers, and of consumers of data flows. We can represent these relationships with three matrices:

- ✓ the *database/organization matrix*, where, for the most relevant databases, we represent organizations that create data and organizations that use data. This matrix could be refined, representing single entities (or tables), but in order to make its size reasonable, we set the granularity at the database level;
- ✓ the *dataflow/organization matrix*, similar to the previous one, in which we represent organizations which provide or consume the most relevant data flows; and
- ✓ the *database/process matrix* in which we represent the processes which provide or consume the most relevant data.

Phase 2: DQ Risk Identification

The *Loss event profiling and evaluation of economic losses* step is in charge of providing for loss events caused by low data quality, the economic value of the expected loss. Values can be: (i) absolutes (e.g., 350 euro), (ii) a percentage with respect to some reference variables (e.g., 4% of turnover), or (iii) a qualitative evaluation (e.g., low or high).

The evaluation of economic losses caused by low data quality, is realized by considering the hierarchy of costs shown in Figure 2. In particular, we have to select items among the costs caused by low data quality. For each significant item, an appropriate metric has to be defined and the corresponding economic value is calculated. Thus, for example, the item *Rework cost*, which is a *re-processing work*, can be measured by multiplying the hours of works realized by operators for the hourly cost of operators. Output of this step are the evaluation of economic losses for each loss event and a classification of loss events according with economic losses and frequency of events.

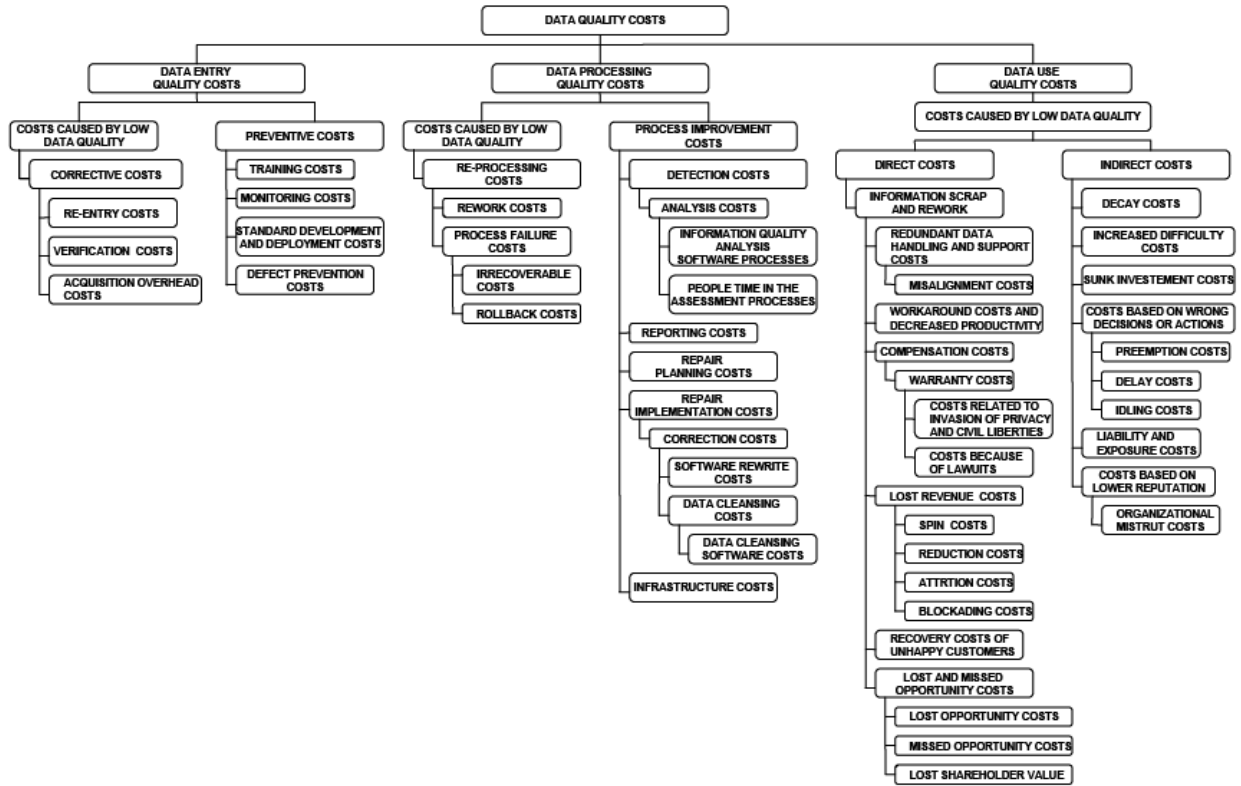


Fig. 2. A comprehensive classification of costs of poor data quality.

The next step in this phase is the *Selection of critical processes*, whose aim is to classify processes according with the process loss value. This value is calculated by considering the economic losses related to event losses caused by low DQ associated to processes for the frequency of the events.

The last step of this phase has the goal of identifying critical data sets to asses.

This activity is performed by considering critical processes defined in the previous step and by identifying the data sets provided or consumed by such processes. Data sets are derived from the data/process matrix created in the state reconstruction phase. The identified data sets and the classification of loss events created before are the rows and columns of a new matrix whose values are quality dimensions related to data set are affected by the specific loss event.

Phase 3: DQ Risk Measurement

After the identification of the relevant data sets to assess and the DQ dimensions along which to realize the assessment, we have to select appropriate metrics for the DQ dimensions. The literature proposed a wide number of metrics both quantitative and qualitative [5]; the data quality expert has to choose the most appropriate metric by considering the feasibility of applying metric to the specific data set. Once the appropriate metric is defined, it is necessary to make or buy software components (probe) able to apply it on the selected data set. Probes results can be correlated to the economic costs of loss event. In order to produce an accurate economic evaluation not only direct costs have to be considered. Other relevant costs are defined in Figure 2. For each of relevant item (e.g. indirect cost such as delay cost) specific metrics are identified.

Phase 4: DQ Risk Monitoring

The last phase of the assessment methodology is focused on the definition of DQ thresholds. When probes register such values alerts can be sent. The definition of thresholds can be realized in many ways. One of the most effective approaches is based on the discriminant analysis [16]. Discriminant analysis is a technique for classifying a set of observations into predefined classes. The purpose is to determine the class of an observation, based on a set of variables known as predictors or input variables. The model is built based on a set of observations (training set) for which the classes are known. In our case we consider two different classes: the class of DQ measurements sessions during which no loss events happened (*NoLoss* class) and the class of DQ measurements sessions during which at least one loss event happened (*Loss* class). Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions, such that

$$L = b_1x_1 + b_2x_2 + \dots + b_nx_n + c$$

where the b 's are discriminant coefficients, the x 's are the input variables or predictors and c is a constant.

The discriminant functions are used to predict for a new observation the class belongs to. For a k class problem k discriminant functions are constructed. Thus when a new DQ measurement is realized by probes it can be assigned to the Loss or NoLoss class.

Architectural Framework

The architecture of the framework is shown in Figure 3. It is composed by five modules (and associated repositories), each one aims to support a specific phase of the methodology. The *Knowledge Extractor* module supports the *DQ Risk prioritization* and the *DQ Risk Identification* phases and it allows: (i) defining all the relationships between data used in processes to produce services realized by organization units and organization units, processes, services, and (ii) storing them in the *Knowledge Repository*. The *Data Quality Assessment* module is in charge of assessing the quality of data sources applying different algorithms and techniques; it supports the *DQ Risk Measurement* phase storing all the results in the *Data Quality Dimension Repository*. The *Analysis* module aims of process information retrieved by probes and calculated by the data quality assessment module supporting *DQ Risk Measurement* phase; the OLAP analysis are stored in the *Analysis Repository*. Finally the *Monitoring & Reporting module*, which supports the *DQ Risk Monitoring* phase allows performing monitoring and reporting activities on the most important information; it uses data stored in the *Monitoring & Reporting Repository*. Finally the *Probe Management Layer* allows configuring and handling probes which are installed in the target database, the probes are in charge of extracting all the data and information used by other modules; all the information regarding probes and their configuration are contained in the *Probes Repository*. The prototype version of the framework is a Web application, based on the Model View Control design pattern and it is developed on the top of Pentaho (www.pentaho.com). The Pentaho BI Project provides enterprise-class reporting, analysis, dashboard, data mining and workflow capabilities that help organizations operate more efficiently and effectively. The framework has been extended to include a dedicated software R project (www.r-project.org) for supporting statistical computation in the analysis module.

In the following sections we describe in more detail each module of the architecture, and we show results of application of our tool to the assessment activity of data in an Italian Public Administration. For privacy reason we cannot show details about internal organization and data structure, but we report the most significant result of such analysis.

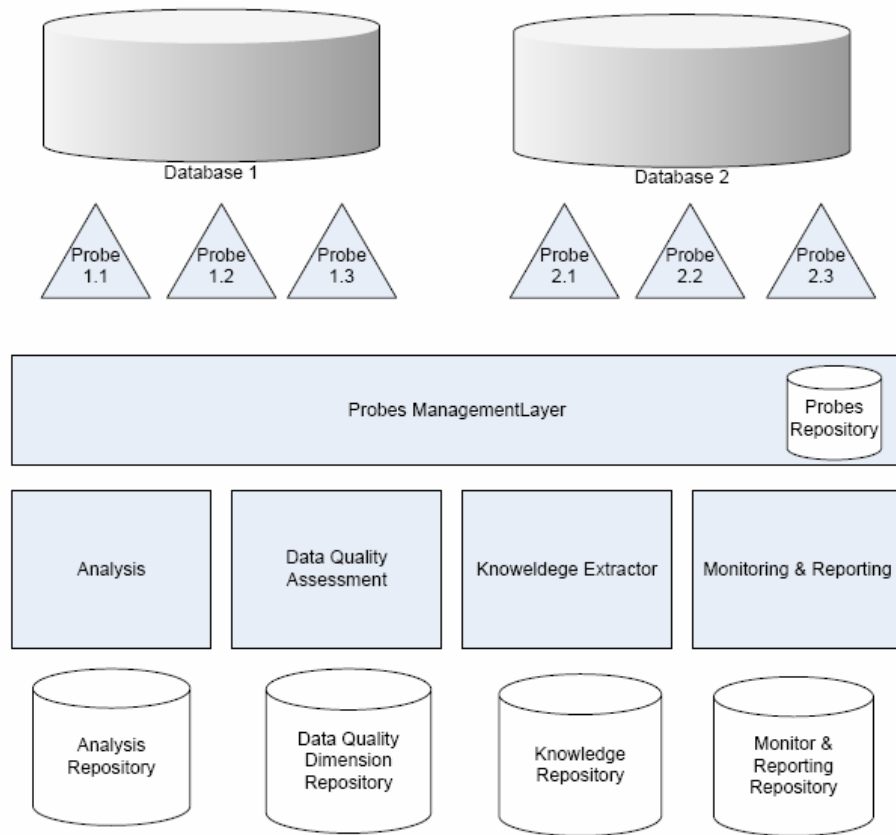


Fig. 3. Tool Architecture.

Knowledge Extractor Module

The *Knowledge Extractor* module supports the phase 1 and 2 of the methodology. It allows data quality expert to extract from sources the information needed to feed the *database/organization matrix*, the *database/process matrix* and the *dataflow/organization matrix*. Matrices are stored in the *Knowledge Repository*, together with all the information regarding the target database, the organizational units, and the data flows. Figure 4 shows a simplify version of the entity relationship schema of the *Knowledge Repository* that it is composed by 35 entities, each one composed of about 10 attributes.

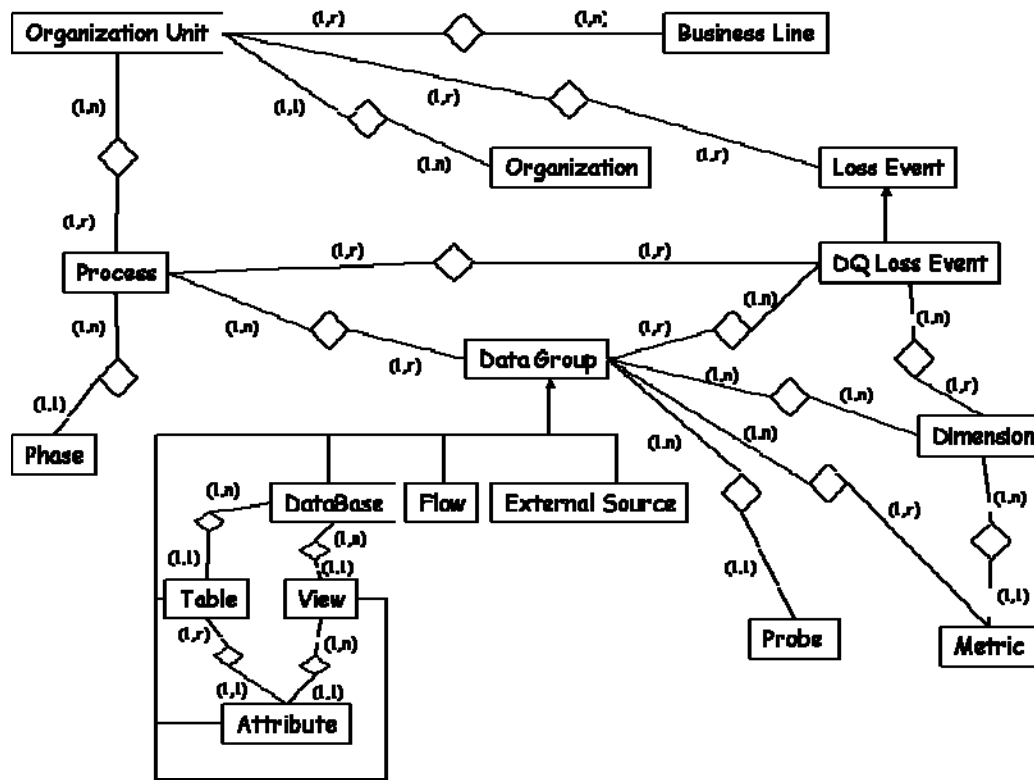


Fig. 4 Simplify version of the entity relationship schema of the Knowledge Repository.

According with Figure 4, an organization unit produces or suffers loss events and in particular loss event related to low data quality. These events are related to specific data quality dimension that can be evaluated by using specific metric. Organization unit are involved in processes that are structured in phases. Processes can produce or suffer DQ loss events. Processes manage Data group representing relational databases, data flows and external sources. Data quality values are measured by means of probes.

Data Quality Assessment Module

The *Data Quality Assessment* module applies data quality techniques and algorithms on data retrieved from the probes for measuring data quality dimensions. The prototype version of the framework supports the following metrics:

- ✓ For the *Syntax Accuracy* dimension, the metric is based on comparison functions which exploit distance of data values, in particular we implement the Jaro and Soundex Code distance metric [12,5]. *Jaro* distance is a string comparison function that define the number of insertions, deletions, and transpositions between two strings. Jaro's algorithm finds the number of common characters and the number of transposed characters in the two strings. A *common character* is a character that appears in both strings within a distance of half the length of the shorter string. A *transposed character* is a common character that appears in different positions. As an example, comparing *Smith* and *Simth*, there are five common characters, two of which are transposed. Instead, the purpose of *Soundex Code* distance is to cluster together names that have similar sounds. For example, the soundex code of *Hilbert* and *Heilbpr* is similar. A soundex code always

contains four characters. The first letter of the name becomes the first character of the soundex code. The remaining three characters are drawn from the name sequentially, by accessing a predefined table. As an example, the soundex code of *Hilbert* and *Heilbpr* is H416. Once the four-character limit has been reached, all remaining letters are ignored.

Jaro which is a character-based similarity comparison works well for typographically errors, instead Soundex works efficiently with phonetically similar string, thus we use the former with attribute values affected by typos, while we use the latter with those attributes storing values with phonetic errors. The prototype version exploits the Jaro and Soundex distance functions implemented in open source libraries available online. In particular, we used *Secondstring* library (available at <http://www.cs.wlu.edu/~levy/kd/>) and *SimMetrics* library (available at <http://sourceforge.net/projects/simmetrics/>).

- ✓ For the *Completeness* dimension, the metric is based on the *closed world assumption* and the relational model with NULL values. As a consequence the metric calculate the ratio of null values on table values
- ✓ For the *Currency* dimension we consider temporal metadata such as *TIMESTAMP* data type provided by the target dbms.
- ✓ For the *Consistency* dimension we adopt two different metrics. The first one is based on record linkage techniques [12] and it is used to verify foreign key consistency rules in presence of dirty data. Another metric is used to verify simple business rules, which have the form of IF – THEN rules. As an example, we can count all the tuples in a table satisfying the following rule: *if marital status is married, age must not be less than 14*, where *marital status* and *age* are attributes, while *married* and *14* are possible values.

The information about dimensions, metrics and results of assessment are stored in the *Data Quality Assessment Repository*. Figure 5 shows the entity relationship schema we adopt to store measurements performed by Web Services probes and the Data Quality Assessment module. According with Figure 5 a *Probe* is described by the *Identifier*, its *Description Name*, the *Description* of probe's features, the *URL* where the Web service can be invoked, the *WSDL* file describing operations performed by the probe and, finally, the *Owner* of the probe, in this way it is possible to manage existing commercial probes that can be already installed in the organization where the assessment activity is realized. A probe is able to perform assessment activities over a set of Data Groups. A *Data Group* is the root of a hierarchy representing a generic relational data structure (or part of it) that can be evaluated. For example a probe can assess one or more *Attribute* (s), *Table* (s), *View* (s) or one or more *Database* (s) composed by tables and views. Each Web service probe provides one or more different *Measurement Type*(s), each one realized by a specific *Metric* related to a given *DQ Dimension*. Each measurement type is described by a *Measurement Unit* and a *Precision*.

Probe executes a number of *Measurements*, each one characterized by an identifier (*IdMeasurement*), the measurement *Value*, and a *Timestamp*. Measurements can be clustered into sets of measurement representing the result of an evaluation campaign for a specific task (e.g. measure the accuracy dimension of a table, every first day of all months). Sets are represented by the *Measurement Family* entity, described by an id (*IdFamily*) and a general *Description*. Finally, each measurement is obtained over one or more instances of the *DataGroup* entity.

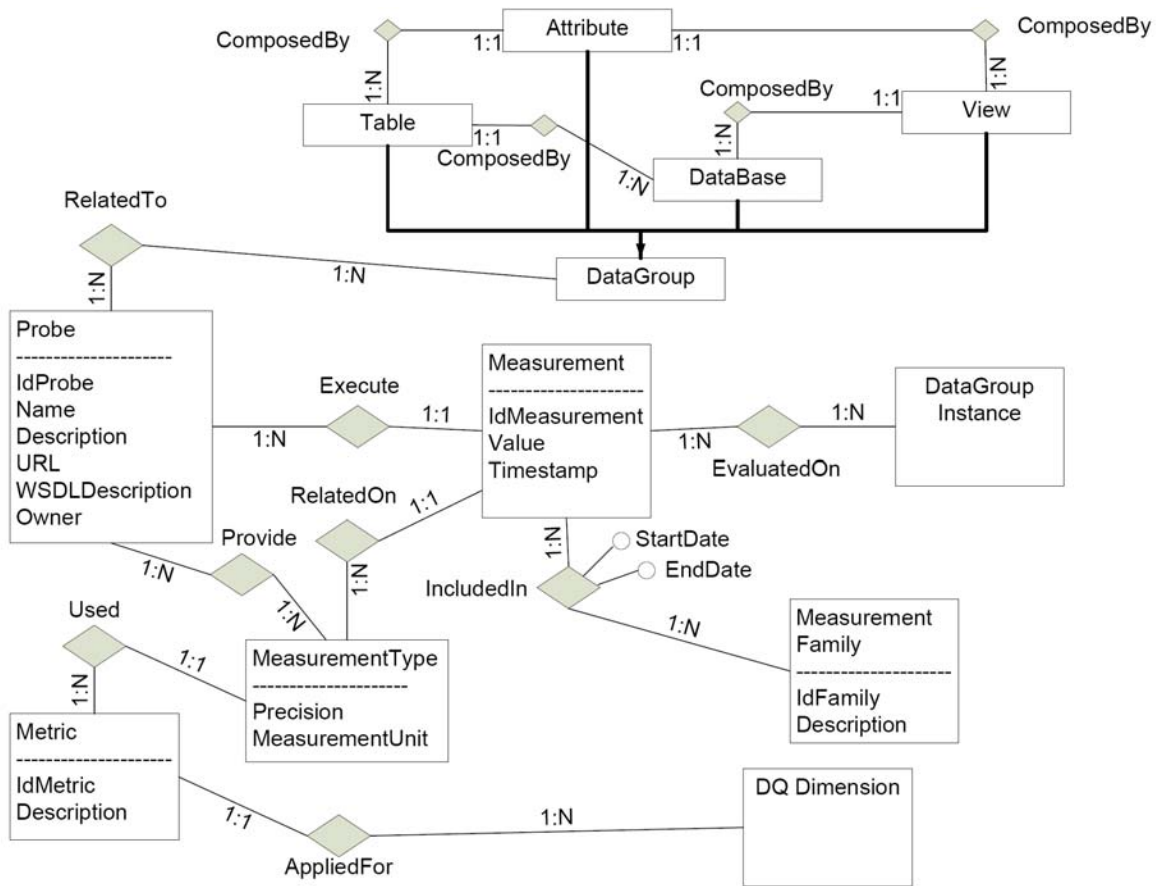


Fig. 5. Entity relationship schema for data assessment.

Analysis Module

The *Analysis* module allows processing information retrieved by probes and calculated by the data quality assessment module. It allows defining and handling aggregation measures, data marts and datawarehouses. Data are organized in cubes (or *hypercubes*), which are defined over a multidimensional space, consisting of several dimensions. Two are data models used to store assessment information, that are strictly related to the physical model supported by existing OLAP solutions. Whereas some vendors, especially vendors of traditional relational database systems (RDBMS), propose the ROLAP architecture (Relational On-Line Analytical Processing), others support the MOLAP architecture (Multidimensional On-Line Analytical Processing) [21]. The advantage of the MOLAP architecture is that it provides a direct multidimensional view of the data whereas the ROLAP architecture is just a multidimensional interface to relational data. On the other hand, the ROLAP architecture has two advantages: (a) it can be easily integrated into other existing relational database systems, and (b) relational data can be stored more efficiently than multidimensional data. For the above described reasons we adopt a ROLAP solution and consequently we defined a star-schema to store all relevant information related to the assessment phase.

The star-schema is stored in the *Analysis Repository* using a ROLAP multidimensional model, where the fact table is the measurement and the dimension tables are among others, time, DQ dimensions, loss events, databases, economic value of loss events.

Starting from this schema, it is possible to produce several kinds of reports according with considered dimension that are used by the *Monitoring and Reporting* module.

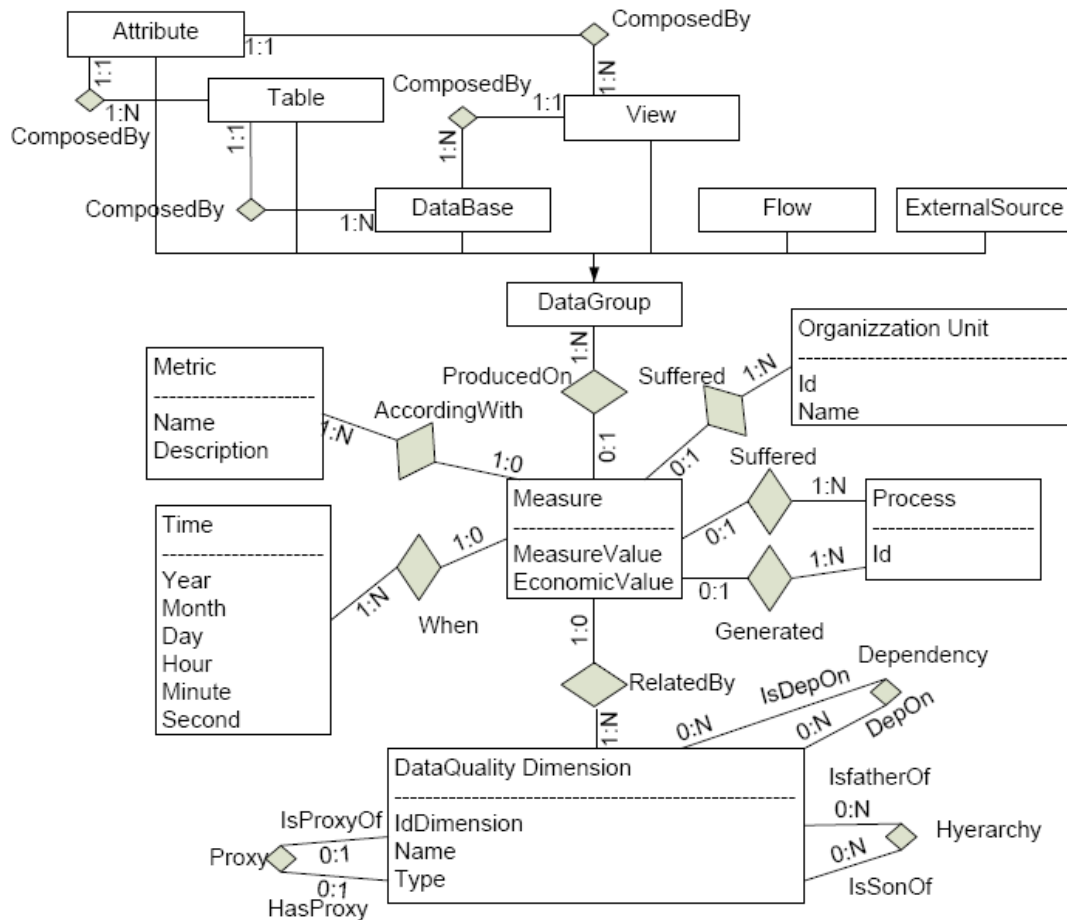


Fig. 6. Star schema used by Analysis Module.

Figure 6 shows the star-schema we developed. The fact table is the *Measure* which is described by the *Measure value*. Six are the dimension tables; namely:

1. *Metric*, which is in charge of describing the specific metric applied to the measure.
2. *Time*, which describes when the measurement was realized.
3. *Data Quality (DQ) Dimension*, which is in charge of reporting the measure to the DQ dimension associated to the measure. The table is enriched with three ring relationships to better describe DQ dimensions. In fact, DQ dimensions may have one or more dependencies with other DQ dimensions [1]. The description of this information, through the *Dependency* relationship, helps the data quality expert to identify causes of low data quality values. Moreover sometimes is very difficult to define a metric for DQ dimensions, in this case it is possible to select a "proxy" dimension which provides an indirect, but easier to measure, evaluation related to the DQ dimension. Proxy dimensions are represented by the *Proxy* relationship. Finally, it is possible to organize DQ dimensions in one or more hierarchies (see for example the ones proposed by [15,7]) through the *Hierarchy* relationship.
4. *Process*, which allows connecting the measure of data with the process that is in charge of creating it.

5. *Organization Unit*, which associates the measure of data with the organizational unit that creates or uses the assessed data.
6. *DataGroup*, which represents a generic relational data structure (or part of it) that can be evaluated. For example a measurement can be produced by assessing one or more *Attributes*, *Tables*, *Views* or one or more *Databases* that are composed by tables and views.

Monitoring & Reporting Module

The *Monitoring & Reporting* is in charge of monitoring data quality assessment results by offering a web-based support. It allows specifying the data quality dimensions to evaluate, the target database and the time intervals of the monitoring activity. It allows producing reports about information of interest. Reports can be realized in different formats such as a pdf file, html page and Microsoft Excel file. The information for the monitoring and reporting activities is stored in the *Monitoring & Reporting Repository*. Moreover the tool is able to send alert messages when assessment values are close to thresholds calculated by applying the discriminant analysis.

In order to monitor the results of the assessment phase we adopt a ROLAP approach, which provides a multidimensional analysis to the underlying information. To achieve this goal, the module uses the star schema defined in the *Analysis* module illustrated in previous subsection. In Figure 7 we provide an example of analysis and monitoring activities related to an Italian Public Administration: the figure shows the monitoring of the number of inaccurate records over a period of four years before and after improvement activities.

It is possible to observe that the number of inaccurate records is already reduced of the 45% from the first year. The projection indicates that in the range of 4 years the number of inaccurate records is reduced to negligible amount.



Fig. 7 Monitoring of inaccurate records in a real case of an Italian Public Administration before and after improvement activities.

Figure 8 shows an example of excel reporting using the *monitoring & reporting module*. The report shows the distribution of errors affecting data belonging to the above Italian Public Administration analyzed in the 2006 year of assessment. It is possible to observe that the main inaccurate values are present in the *address* and *value added tax number* attributes, while other inaccurate values are negligible.

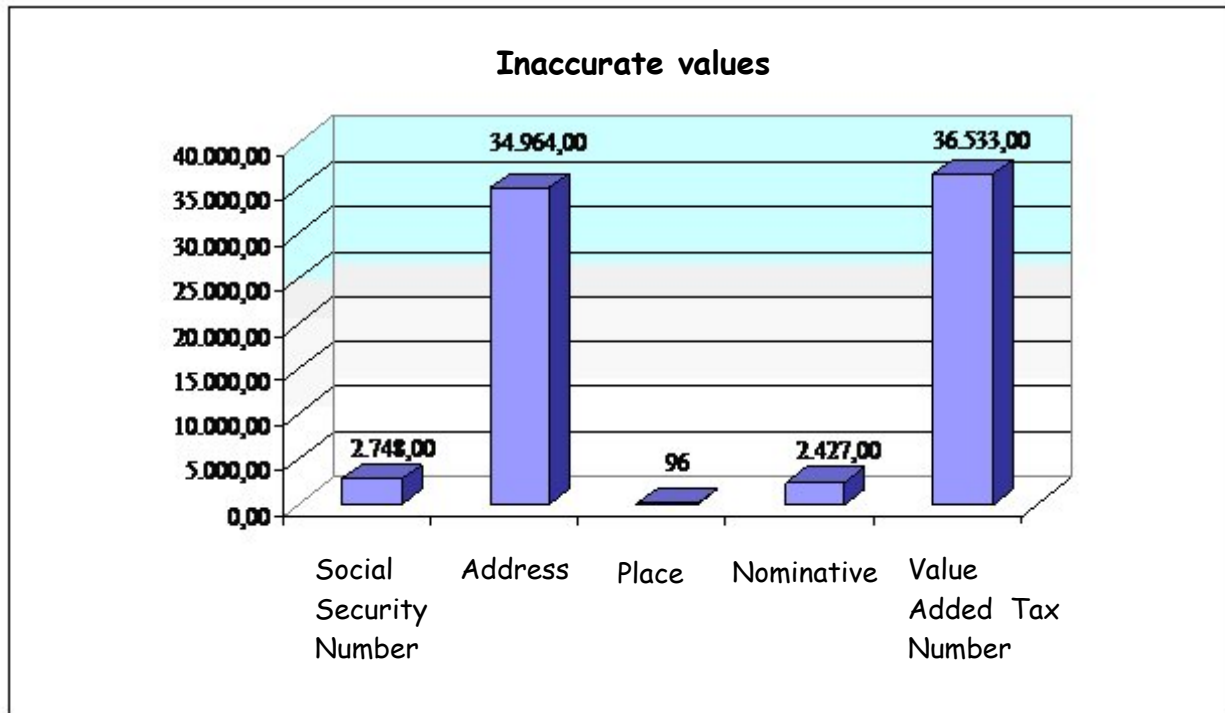


Fig. 8 Distribution of errors affecting data belonging to an Italian Public Administration.

Probe Management Layer Module

The *Probe Management Layer* allows configuring and handling probes which are installed in the target database. A probe is a software component, exposed as Web service, which is in charge of extracting from the database needed information that will be used to evaluate a specific data quality dimension. For example in order to measure the completeness of an attribute of a table, a probe extracts all values of the attribute. Each probe may be used for different measurements. The *Probe Management Layer* allows configuring different parameters of a probe. For example a parameter can specify if the probe retrieves information about the whole database or about a single table or about a single attribute. Parameters about each probe are stored in the *Probes Repository* not shown here for the sake of brevity.

Related Work

In the literature, several are the proposals of assessment methodologies. Table 1 analyzes 14 different methodologies (including ORME-DQ) [20, 11, 6, 13, 14, 17, 8, 9, 19, 15, 18, 2, 5] compared with respect to the typical steps of the assessment phase [4]. The most commonly addressed steps of the assessment phase are *data analysis* and *measurement of quality*. However, they are performed according to different approaches. For example, the *measurement of quality* step is performed with questionnaires in AIMQ, with a combination of subjective and objective metrics in DQA, or with statistical analyses in QAFD. In general, subjective metrics seem more suitable for exploratory analyses emphasizing users' judgments. Statistical measures of quality, such as average completeness or maximum accuracy, are often applied to data used in decision-making processes. However, in most cases, the variety of contextual requirements involves the use of a mix of subjective, objective or statistical measures of quality, consistent with the

breadth of DQ metrics. Only a few methodologies consider the *DQ requirements analysis*, identifying DQ issues and collecting target DQ levels of users. This step is particularly relevant to evaluate and resolve conflicts in target DQ levels from different stakeholders. For example, QAFD recommends the collection of target DQ levels from different types of experts, including business experts and financial operators, but does not provides guidelines for the reconciliation of incompatible DQ levels. ORME-DQ is the only methodology that adopts a probabilistic approach for identifying the new DQ levels. The last column of Table 1 specifies whether the methodology allows extensibility to other metrics than those explicitly dealt with in the methodology. E.g. ORME-DQ mentions specific dimensions and metrics, but, in the description of the steps, the approach followed is easily generalizable to other dimensions and metrics. On the contrary, ISTAT provides detailed measurement and improvement procedures for accuracy, completeness and consistency, so that the whole approach is strictly "hardwired" to such dimensions. The interested reader can refer to [4].

Step/ Methodologies Acronym	Data Analysis	DQ Requirement Analysis	Identification of Critical Areas	Process Modeling	Measurement of quality	Extensible to other metrics
TDQM [20]	+		+	+	+	Fixed
DWQ [12]	+	+	+		+	Open
TQDM [6]	+	+	+	+	+	Fixed
AIMQ [13]	+		+		+	Fixed
CIHI [11]	+		+			Fixed
DQA [17]	+		+		+	Open
IQM [8]	+				+	Open
ISTAT [9]	+				+	Fixed
AMEQ [10]	+		+	+	+	Open
COLDQ [15]	+	+	+	+	+	Open
DaQuinCIS [18]	+		+	+	+	Open
QAFD [2]	+	+			+	Fixed
CDQ [4]	+	+	+	+	+	Open
ORME-DQ	+		+	+	+	Open

Table 1. Comparison of methodologies and assessment steps

Conclusion

In this paper we presented ORME-DQ, a methodology and a framework for data quality assessment. The methodology has been developed following the Basel II approach on operational risk and it is based on the notion of even loss caused by low data quality. The prototype version of the framework has been realized on the top of Pentaho software and it is extended with specific tools for data analysis. We are currently working on: (i) the extension of the ORME-DQ by adding the improvement steps and (ii) the definition of algorithms for the identification of the amount of capital that has to be allocated in order to considered economic losses derived by low data quality as foreseen in the Basel II agreements.

Acknowledgments

The work presented in this paper has been partially supported by the Italian FIRB project RBNE05XYPW NeP4B - Networked Peers for Business, the Italian FIRB project eG4M eGovernment for Mediterranean

countries, and the Italian ORME project.

REFERENCES

- [1] F. D. Amicis, D. Barone, and C. Batini. An analytical framework to analyze dependencies among data quality dimensions. In Proceedings of the Eleven International Conference on Information Quality 2006 (ICIQ06) Boston, Massachusetts, USA. MIT press, 2006.
- [2] F. D. Amicis and C. Batini. A methodology for data quality assessment on financial data. *Studies in Communication Sciences*, 2004.
- [3] Basel. Basel II: International convergence of capital measurement and capital standards: a revised framework. Technical report, 2004.
- [4] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. Submitted to *Computing Surveys*, 2007.
- [5] C. Batini and M. Scannapieco. *Data Quality. Concepts, Methods and Techniques*. Springer Verlag, 2006.
- [6] L. English. *Improving Data Warehouse and Business Information Quality*. Wiley & Sons, 1999.
- [7] M. Eppler and M. Helfert. A classification and analysis of data quality costs. In Proceedings of the Ninth International Conference on Information Systems (ICIQ- 04), Boston, 2004.
- [8] M. Eppler and P. Muenzenmaier. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In Proceedings of the Seventh International Conference on Information Systems (ICIQ-04), Boston, 2002.
- [9] P. Falorsi, S. Pallara, A. Pavone, A. Alessandrini, E. Massella, and M. Scannapieco. Improving the Quality of Toponymic Data in the Italian Public Administration. In Proceedings of the ICDT 03 Workshop on Data Quality in Cooperative Information Systems (DQCIS '03), Siena, Italy, 2003.
- [10] D. W. Institute. Data quality and the bottom line: Achieving business success through a commitment to high quality data. <http://www.dw-institute.com/>.
- [11] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, editors. *Fundamentals of Data Warehouses*. Springer Verlag, 1995.
- [12] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In SIGMOD Conference, pages 802–803, 2006.
- [13] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. Aimq: A methodology for information quality assessment. *Information and Management*, 2001.
- [14] J. Long and C. Seko. A Cyclic-Hierarchical Method for Database Data-Quality Evaluation and Improvement. In R. Wang, E. Pierce, S. Madnick, and F. C.W., editors, *Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph*. Sharpe, M.E., April 2005.
- [15] D. Loshin. *Enterprise Knowledge Management - The Data Quality Approach - Chapter 4*. Morgan Kaufmann Series in Data Management Systems, 2004.
- [16] G. McLachlan, E. Kuh, and R. Welsh. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience, 2004.
- [17] L. Pipino, Y. Lee, and R. Wang. Data Quality Assessment. *Communications of the ACM*, 45(4), 2002.
- [18] M. Scannapieco, A. Virgillito, M. Marchetti, M. Mecella, and R. Baldoni. The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. *Information Systems*, 29(7), 2004.
- [19] Y. Su and Z. Jin. A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In Proceedings of the Ninth International Conference on Information Quality 2004 (ICIQ04), pages 447–465, Boston, MA, USA, 2004.
- [20] R. Wang. A Product Perspective on Total Data Quality Management. *Comm. of the ACM*, 41(2), 1998.
- [21] The OLAP Survey 6, technical report, <http://www.olapreport.com/survey.htm>, 2007.