Methodologies for Data Quality Assessment and Improvement

CARLO BATINI

Università di Milano - Bicocca

CINZIA CAPPIELLO

Politecnico di Milano

CHIARA FRANCALANCI

Politecnico di Milano

and

ANDREA MAURINO

Università di Milano - Bicocca

The literature provides a wide range of techniques to assess and improve the quality of data. Due to the diversity and complexity of these techniques, research has recently focused on defining methodologies that help the selection, customization, and application of data quality assessment and improvement techniques. The goal of this article is to provide a systematic and comparative description of such methodologies. Methodologies are compared along several dimensions, including the methodological phases and steps, the strategies and techniques, the data quality dimensions, the types of data, and, finally, the types of information systems addressed by each methodology. The article concludes with a summary description of each methodology.

Categories and Subject Descriptors: A.1 [**Introductory and Survey**]; H.2.m [**Database Management**]: Miscellaneous

General Terms: Management, Measurement

Additional Key Words and Phrases: Data quality, data quality measurement, data quality assessment, data quality improvement, methodology, information system, quality dimension

ACM Reference Format:

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41, 3, Article 16 (July 2009), 52 pages.
DOI = 10.1145/1541880.1541883 http://doi.acm.org/10.1145/1541880.1541883

This article has been partially supported by the European IST project n. 27347 SEEMP - Single European Employment Market-Place, the European IST project WSDiamond - Web Services DIAgnosability, Monitoring and Diagnosis and the Italian FIRB project Eg4M - eGovernment for Meditteranean countries.

Author's address: C. Batini, Dipartimento di Informatica, Sistemistica e Comunicazioni (DISCO), Universita' degli Studi di Milano Bicocca, Viale Sarca 336, 20126, Milano, Italy; email: batini@disco.unimib.it

DOI 10.1145/1541880.1541883 http://doi.acm.org/10.1145/1541880.1541883

1. INTRODUCTION TO DATA QUALITY

Because electronic data are so pervasive, the quality of data plays a critical role in all business and governmental applications. The quality of data is recognized as a relevant performance issue of operating processes [Data Warehousing Institute 2006], of decision-making activities [Chengalur-Smith et al. 1999], and of interorganizational cooperation requirements [Batini and Scannapieco 2006]. Several initiatives have been launched in the public and private sectors, with data quality having a leading role, such as the Data Quality Act enacted by the United States government in 2002 [Office of Management and Budget 2006] and the Data Quality Initiative Framework enacted by the government of Wales in 2004 to improve the information quality of all general medical practices [DQI 2004].

At the same time, information systems have been migrating from a hierarchical/monolithic to a network-based structure, where the set of potential data sources that organizations can use has dramatically increased in size and scope. The issue of data quality has become more complex and controversial as a consequence of this evolution. In networked information systems, processes are involved in complex information exchanges and often operate on input obtained from external sources, which are frequently unknown a priori. As a consequence, the overall quality of the data that flows across information systems can rapidly degrade over time if the quality of both processes and information inputs is not controlled. On the other hand, networked information systems offer new opportunities for data quality management, including the availability of a broader range of data sources and the ability to select and compare data from different sources to detect and correct errors, and, thus, improve the overall quality of data.

The literature provides a wide range of techniques to assess and improve the quality of data, such as record linkage, business rules, and similarity measures. Over time, these techniques have evolved to cope with the increasing complexity of data quality in networked information systems. Due to the diversity and complexity of these techniques, research has recently focused on defining methodologies that help select, customize, and apply data quality assessment and improvement techniques. This article defines a *data quality methodology* as a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data. The goal of this article is to provide a systematic and comparative description of existing data quality methodologies.

The article is organized as follows. Section 2 introduces the basic data quality issues common to all methodologies, which represent the perspectives used in this article for comparative analysis such as: (1) the methodological phases and steps, (2) the strategies and techniques, (3) the data quality dimensions, (4) the types of data, and, finally, (5) the types of information systems. Section 3, the core of the article, compares existing methodologies along the coordinates introduced in Section 2. The comparison is performed with: synoptic tables, that highlight at a glance groups of methodologies with similar approaches, in-depth comments, and qualitative evaluations. Section 4 describes ongoing research and future research directions in the field of data quality methodologies. Finally, the article concludes with a summary description of each methodology in Appendix A describing: (1) the phases of each methodology and their mutual dependencies and critical decisions, (2) a general description highlighting the focus of each methodology and original contribution to the data quality assessment and improvement process, and (3) detailed comments discussing the applicability of each methodology.

2. COMPARATIVE PERSPECTIVES FOR ANALYZING METHODOLOGIES

There exist several perspectives that can be used to analyze and compare data quality (DQ) methodologies:

(1) the *phases and steps* that compose the methodology;
(2) the *strategies and techniques* that are adopted in the methodology for assessing and improving data quality levels;
(3) the *dimensions and metrics* that are chosen in the methodology to assess data quality levels;
(4) the *types of costs* that are associated with data quality issues including:
   (a) *costs associated with poor data quality*, that is process costs caused by data errors and opportunity costs due to lost and missed revenues; these costs are also referred to as *indirect costs*;
   (b) *costs of assessment and improvement* activities, also referred as *direct costs*;
(5) the *types of data* that are considered in the methodology;
(6) the *types of information systems* that use, modify, and manage the data that are considered in the methodology;
(7) the *organizations* involved in the processes that create or update the data that are considered in the methodology, with their structure and norms;
(8) the *processes* that create or update data with the goal of producing services required by users that are considered in the methodology;
(9) the *services* that are produced by the processes that are considered in the methodology.

Methodologies differ in how they consider all of these perspectives. In the remainder of this article the last three perspectives—organization, process, and service—will not be investigated, as they are rarely mentioned in methodologies.

2.1. Common Phases and Steps

In the most general case, the sequence of activities of a data quality methodology is composed of three phases:

(1) *State reconstruction*, which is aimed at collecting contextual information on organizational processes and services, data collections and related management procedures, quality issues and corresponding costs; this phase can be skipped if contextual information is available from previous analyses.
(2) *Assessment/measurement*, which measures the quality of data collections along relevant quality dimensions; the term *measurement* is used to address the issue of measuring the value of a set of data quality dimensions. The term *assessment* is used when such measurements are compared to reference values, in order to enable a diagnosis of quality. The term assessment is adopted in this article, consistent with the majority of methodologies, which stress the importance of the causes of poor data quality.
(3) *Improvement* concerns the selection of the steps, strategies, and techniques for reaching new data quality targets.

The state reconstruction phase is optional if the assessment phase can be based on existing documentation. Since methodologies typically make this assumption, we will not further discuss the state reconstruction phase. Although adopting different names, methodologies organize the assessment and improvement phases in terms of a common set of basic *steps*. The steps of the assessment phase are:

—*data analysis*, which examines data schemas and performs interviews to reach a complete understanding of data and related architectural and management rules;

—*DQ requirements analysis*, which surveys the opinion of data users and administrators to identify quality issues and set new quality targets;

—*identification of critical areas*, which selects the most relevant databases and data flows to be assessed quantitatively;

—*process modeling*, which provides a model of the processes producing or updating data;

—*measurement of quality*, which selects the quality dimensions affected by the quality issues identified in the DQ requirements analysis step and defines corresponding metrics; measurement can be *objective* when it is based on quantitative metrics, or *subjective*, when it is based on qualitative evaluations by data administrators and users.

Note that in all the steps of the assessment phase, a relevant role is played by *metadata* that store complementary information on data for a variety of purposes, including data quality. Metadata often provide the information necessary to understand data and/or evaluate them.

The steps of the improvement phase are:

—*evaluation of costs*, which estimates the direct and indirect costs of data quality;

—*assignment of process responsibilities*, which identifies the process owners and defines their responsibilities on data production and management activities;

—*assignment of data responsibilities*, which identifies the data owners and defines their data management responsibilities;

—*identification of the causes of errors*, which identifies the causes of quality problems;

—*selection of strategies and techniques*, which identifies all the data improvement strategies and corresponding techniques, that comply with contextual knowledge, quality objectives, and budget constraints;

—*design of data improvement solutions*, which selects the most effective and efficient strategy and related set of techniques and tools to improve data quality;

—*process control*, which defines check points in the data production processes, to monitor quality during process execution;

—*process redesign*, which defines the process improvement actions that can deliver corresponding DQ improvements;

—*improvement management*, which defines new organizational rules for data quality;

—*improvement monitoring*, which establishes periodic monitoring activities that provide feedback on the results of the improvement process and enables its dynamic tuning.

In Section 3.1, methodologies are compared in their assessment and improvement capabilities by evaluating their completeness along the set of phases and steps introduced in this section. Note that, usually, each methodology refers to a specific assessment or improvement functionality by using different terms. In the appendix, we describe methodologies by adopting the original terms, but we provide the correspondence between such terms and the classification presented here.

2.2. Strategies and Techniques

In their improvement steps, methodologies adopt two general types of *strategies*, namely *data-driven* and *process-driven*. *Data-driven* strategies improve the quality of data by directly modifying the value of data. For example, obsolete data values are updated by refreshing a database with data from a more current database. *Process-driven* strategies improve quality by redesigning the processes that create or modify data. As an example, a process can be redesigned by including an activity that controls the format of data before storage.

Strategies, both data- and process-driven, apply a variety of techniques: algorithms, heuristics, and knowledge-based activities, whose goal is to improve data quality. An open-ended list of the improvement techniques applied by data-driven strategies is:

(1) *acquisition of new data*, which improves data by acquiring higher-quality data to replace the values that raise quality problems;

(2) *standardization (or normalization)*, which replaces or complements nonstandard data values with corresponding values that comply with the standard. For example, nicknames are replaced with corresponding names, for example, Bob with Robert, and abbreviations are replaced with corresponding full names, for example, Channel Str. with Channel Street.

(3) *Record linkage*, which identifies that data representations in two (or multiple) tables that might refer to the same real-world object;

(4) *data and schema integration*, which define a unified view of the data provided by heterogeneous data sources. Integration has the main purpose of allowing a user to access the data stored by heterogeneous data sources through a unified view of these data. In distributed, cooperative, and P2P information systems (see Section 2.6), data sources are characterized by various kinds of heterogeneities that can be generally classified into (1) technological heterogeneities, (2) schema heterogeneities, and (3) instance-level heterogeneities. *Technological heterogeneities* are due to the use of products by different vendors, employed at various layers of an information and communication infrastructure. *Schema heterogeneities* are primarily caused by the use of (1) different data models, as in the case of a source that adopts the relational data model and a different source that adopts the XML data model, and (2) different representations for the same object, such as two relational sources that represent an object as a table and an attribute. *Instance-level heterogeneities* are caused by different, conflicting data values provided by distinct sources for the same objects. For instance, this type of heterogeneity can be caused by independent and poorly coordinated processes that feed the different data sources. Data integration must face all the types of these listed heterogeneities.

(5) *Source trustworthiness*, which selects data sources on the basis of the quality of their data;

(6) *error localization and correction*, which identify and eliminate data quality errors by detecting the records that do not satisfy a given set of quality rules. These techniques are mainly studied in the statistical domain. Compared to elementary data, aggregate statistical data, such as average, sum, max, and so forth are less sensitive to possibly erroneous probabilistic localization and correction of values. Techniques for error localization and correction have been proposed for inconsistencies, incomplete data, and outliers [Dasu and Johnson 2003]; [Batini and Scannapieco 2006].

(7) *Cost optimization*, defines quality improvement actions along a set of dimensions by minimizing costs.

Two main techniques characterize process-driven strategies:

—*Process control* inserts checks and control procedures in the data production process when: (1) new data are created, (2) data sets are updated, or (3) new data sets are accessed by the process. In this way, a reactive strategy is applied to data modification events, thus avoiding data degradation and error propagation.

—*Process redesign* redesigns processes in order to remove the causes of poor quality and introduces new activities that produce data of higher quality. If process redesign is radical, this technique is referred to as *business process reengineering* [Hammer and Champy 2001]; [Stoica et al. 2003].

Several techniques typical of data- and process- driven strategies are compared in Redman [1996] by discussing the improvement that each technique can achieve along different quality dimensions and the implementation cost of each technique. This comparison is performed from both a short-term and a long-term perspective. The comparison focuses on: (1) acquisition of new data, (2) record linkage, (3) error localization and correction, (4) process control, and (5) process redesign techniques. In general, in the long term, process-driven techniques are found to outperform data-driven techniques, since they eliminate the root causes of quality problems. However, from a short-term perspective, process redesign can be extremely expensive [Redman 1996][English 1999]. On the contrary, data-driven strategies are reported to be cost efficient in the short term, but expensive in the long term. They are suitable for one-time application and, thus, they are recommended for static data.

2.3. Dimensions

In all methodologies, the definition of the qualities, *dimensions*, and *metrics* to assess data is a critical activity. In general, multiple metrics can be associated with each quality dimension. In some cases, the metric is unique and the theoretical definition of a dimension coincides with the operational definition of the corresponding metric. For this reason, in the following we make a distinction between theoretical and operational definitions of dimensions only when the literature provides multiple metrics.

Quality dimensions can be referred either to the *extension* of data—to data values, or to their *intension*—to their schema. Although the quality of *conceptual* and *logical data schemas* is recognized to be a relevant research area [IWCMQ 2003], most definitions of data quality dimensions and metrics are referred to *data values* as opposed to *schemas*. This article focuses mainly on quality dimensions and metrics referred to data values.

The data quality literature provides a thorough classification of data quality dimensions; however, there are a number of discrepancies in the definition of most dimensions due to the contextual nature of quality. The six most important classifications of quality dimensions are provided by Wand and Wang [1996]; Wang and Strong [1996]; Redman [1996]; Jarke et al. [1995]; Bovee et al. [2001]; and Naumann [2002]. By analyzing these classifications, it is possible to define a basic set of data quality dimensions, including *accuracy*, *completeness*, *consistency*, and *timeliness*, which constitute the focus of the majority of authors [Catarci and Scannapieco 2002].

However, no general agreement exists either on which set of dimensions defines the quality of data, or on the exact meaning of each dimension. The different definitions provided in the literature are discussed in the following.

Accuracy. Several definitions are provided for the term *accuracy*. Wang and Strong [1996] define accuracy as “the extent to which data are correct, reliable and certified.” Ballou and Pazer [1985] specify that data are accurate when the data values stored in the database correspond to real-world values. In Redman [1996], accuracy is defined

Table I. Different Definitions Provided for Completeness

Reference	Definition
Wand and Wang 1996	Ability of an information system to represent every meaningful state of a real world system
Wang and Wand 1996	Extent to which data are of sufficient breadth, depth, and scope for the task at hand
Redman 1996	Degree to which values are included in a data collection
Jarke et al. 1995	Percentage of real-world information entered in data sources and/or data warehouse
Bovee et al. 2001	Information having all required parts of an entity's description
Naumann 2002	Ratio between the number of non-null values in a source and the size of the universal relation
Liu and Chi 2002	All values that are supposed to be collected as per a collection theory

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. Data quality methodologies only consider syntactic accuracy and define it as 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'*.

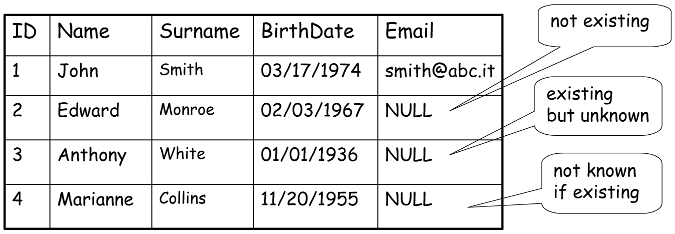
Completeness. Completeness is defined as the degree to which a given data collection includes data describing the corresponding set of real-world objects.

Table I reports the research contributions that provide a definition of completeness. By comparing such definitions, it can be observed that there is a substantial agreement on the abstract definition of completeness. Definitions differ in the context to which they refer, for example, information system in Wand and Wang [1996], data warehouse in Jarke et al. [1995], entity in Bovee et al. [2001].

In the research area of relational databases, completeness is often related to the meaning of *null* values. A null value has the general meaning of *missing value*, 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 Atzeni and Antonellis [1993]). Let us consider the table `Person` reported in Figure 1, with attributes `Name`, `Surname`, `BirthDate`, and `Email`. If the person represented by tuple 2 has no email, tuple 2 is complete. If it is not known whether the person represented by tuple 4 has an email, incompleteness may or may not occur. During quality assessment, a Boolean value (complete or not complete) should be associated with each field to calculate completeness as the ratio between complete values and the total number of values, both at the tuple and at the source level.

Consistency. The consistency dimension refers to the violation of semantic rules defined over a set of data items. With reference to the relational theory, integrity constraints are a type of such semantic rules. In the statistical field, data edits are typical semantic rules that allow for consistency checks.

In the relational theory, two fundamental categories of integrity constraints can be distinguished, namely: *intra-relation constraints* and *inter-relation constraints*. Intra-relation constraints define the range of admissible values for an attribute's domain. Examples are "`Age` must range between 0 and 120," or "If `WorkingYears` is lower than 3, then `Salary` cannot be higher than 25.000 euros per year." Inter-relation integrity



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

not existing
existing but unknown
not known if existing

Fig. 1. Null values and data completeness.

Table II. Existing Definitions of Time-Related Dimensions

Reference	Definition
Wand and Wang 1996	Timeliness refers only to the delay between a change of a real world state and the resulting modification of the information system state
Wang and Wand 1996	Timeliness is the extent to which the age of data is appropriate for the task at hand
Redman 1996	Currency is the degree to which a datum is up-to-date. A datum value is upto- date if it is correct in spite of possible discrepancies caused by timere-lated changes to the correct value
Jarke et al. 1995	Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real world
Bovee et al. 2001	Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long ago it was recorded. Volatility is a measure of information instability, the frequency of change of the value for an entity attribute
Naumann 2002	Timeliness is the average age of the data in a source
Liu and Chi 2002	Timeliness is the extent to which data are sufficiently up-to-date for a task

constraints involve attributes from different relations. As an example, let us consider a `Movies` relation that includes the `Title`, `Director`, and `Year` attributes and an `OscarAwards` relation, specifying the `MovieTitle` and the `Year` when the award was won. An inter-relation constraint could state that for each movie appearing in both relations, "`Movies.Year` must be equal to `OscarAwards.Year`." There is an extensive literature on consistent databases. For example, Arenas et al. [1999], considers the problem of the logical characterization of the notion of *consistent answer* in a relational database, which may violate given integrity constraints. The authors propose a method for computing consistent answers, by proving their soundness and completeness. Integrity constraints have also been studied as enablers of data integration [Calì et al. 2004].

In the statistical area, data from census questionnaires have a structure corresponding to the questionnaire schema. Semantic rules, called *edits*, can be defined on the questionnaire schema to specify the correct set of answers. Such rules typically denote error conditions. For example, an edit could be: if `MaritalStatus` is "married," `Age` must not be lower than 14. After the detection of erroneous records, the act of restoring correct values is called *imputation* [Fellegi and Holt 1976].

Time-related Dimensions: Currency, Volatility, and Timeliness. An important aspect of data is their update over time. The main time-related dimensions proposed in the literature are currency, volatility, and timeliness. Table II compares the definitions provided in the literature for these three time dimensions. Wand and Wang [1996] and Redman

[1996] provide very similar definitions for timeliness and currency. Wang and Strong [1996] and Liu and Chi [2002] assume the same meaning for timeliness, while Bovee et al. [2001] provides a definition for timeliness in terms of currency and volatility. The definition of currency expressed in Bovee et al. [2001] corresponds to timeliness as defined by Wang and Strong [1996] and Liu and Chi [2002]. This comparison shows that there is no agreement on the abstract definition of time-related dimensions; typically, currency and timeliness are often used to refer to the same concept.

2.4. Costs

Costs are a relevant perspective considered in methodologies, due to the effects of low quality data on resource consuming activities. The cost of data quality is the sum of the *cost of data quality assessment and improvement activities*, also referred to as the *cost of the data quality program* and the *cost associated with poor data quality*. The cost of poor quality can be reduced by implementing a more effective data quality program, which is typically more expensive. Therefore, by increasing the cost of the data quality program, the cost of poor data quality is reduced. This reduction can be seen as the benefit of a data quality program.

The cost of a data quality program can be considered a preventive cost that is incurred by organizations to reduce data errors. This cost category includes the cost of all phases and steps that compose a data quality assessment and improvement process (see Section 2.1).

The costs of poor quality can be classified as follows [English 1999]:

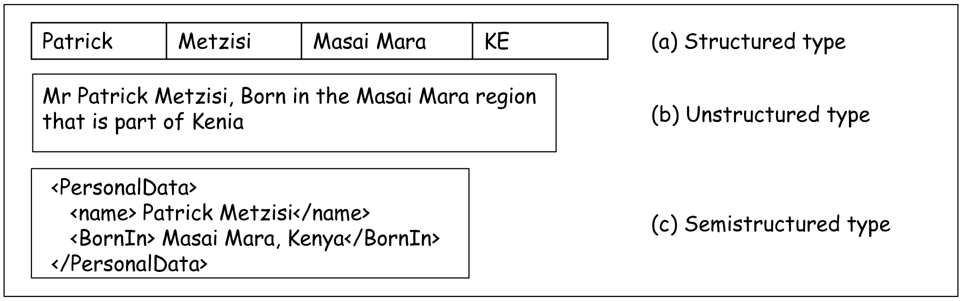
(1) *process costs*, such as the costs associated with the re-execution of the whole process due to data errors;
(2) *opportunity costs* due to lost and missed revenues.

The cost of poor data quality is strongly context-dependent as opposed to the cost of a data quality program. This makes its evaluation particularly difficult, as the same data value and corresponding level of quality has a different impact depending on the recipient. For example, an active trader receiving obsolete information on a stock may incur considerable economic losses as a consequence of wrong investment decisions. In contrast, a newspaper receiving the same obsolete information to publish monthly trading reports may not experience any economic loss.

2.5. Types of Data

The ultimate goal of a DQ methodology is the analysis of data that, in general, describe real world objects in a format that can be stored, retrieved, and processed by a software procedure, and communicated through a network. In the field of data quality, most authors either implicitly or explicitly distinguish three types of data:

(1) *Structured data*, is aggregations or generalizations of items described by elementary attributes defined within a domain. Domains represent the range of values that can be assigned to attributes and usually correspond to elementary data types of programming languages, such as numeric values or text strings. Relational tables and statistical data represent the most common type of structured data.
(2) *Unstructured data*, is a generic sequence of symbols, typically coded in natural language. Typical examples of unstructured data are a questionnaire containing free text answering open questions or the body of an e-mail.
(3) *Semistructured data*, is data that have a structure which has some degree of flexibility. Semistructured data are also referred to as *schemaless* or *self-describing*



Patrick	Metzisi	Masai Mara	KE

Mr Patrick Metzisi, Born in the Masai Mara region that is part of Kenia

```
<PersonalData>
  <name> Patrick Metzisi</name>
  <BornIn> Masai Mara, Kenya</BornIn>
</PersonalData>
```

Fig. 2. Different representations of the same real-world object.

[Abiteboul et al. 2000; Buneman 1997; Calvanese et al. 1999]. XML is the markup language commonly used to represent semistructured data. Some common characteristics are: (1) data can contain fields not known at design time; for instance, an XML file does not have an associated XML schema file; (2) the same kind of data may be represented in multiple ways; for example, a date might be represented by one field or by multiple fields, even within a single data set; and (3) among the fields known at design time, many fields will not have values.

Data quality techniques become increasingly complex as data lose structure. For example, let us consider a registry describing personal information such as `Name`, `Surname`, `Region`, and `StateOfBirth`. Figure 2 shows the representation of Mr. Patrick Metzisi, born in the Masai Mara region in Kenya, by using a structured (Figure 2(a)), unstructured (Figure 2(b)), and semistructured (Figure 2(c)) type of data. The same quality dimension will have different metrics according to the type of data. For instance, syntactic accuracy is measured as described in Section 2.3 in the case of structured data. With semistructured data, the distance function should consider a global distance related to the shape of the XML tree in addition to the local distance of fields.

The large majority of research contributions in the data quality literature focuses on structured and semistructured data. For this reason, although we acknowledge the relevance of unstructured data, this article focuses on structured and semistructured data.

An orthogonal classification of data in the data quality literature is based on viewing data as a manufacturing product [Shankaranarayan et al. 2000]. From this perspective, three types of data are distinguished:

—*raw data items*, defined as data that have not undergone any processing since their creation and first storage—they can be stored for long periods of time;

—*information products*, which are the result of a manufacturing activity performed on data;

—*component data items*, which are generated every time the corresponding information product is required and are stored temporarily until the final product is manufactured.

As will be discussed in Section 3, this classification allows the application to data of quality techniques traditionally used for quality assurance in manufacturing processes.

2.6. Types of Information Systems

DQ methodologies are influenced by the type of information system they refer to both in assessment and in improvement activities. The literature provides the concept of *information system architecture* (IS architecture) to describe the coordination model supported by a company's information system [Zachman 2006]. Different IS architectures or, simply, *types of information systems* are distinguished on the basis of the degree of data, process and management integration supported by a technical system. As the degree of integration of data, process, and management decreases, the data quality assessment and improvement techniques that can be applied become more sophisticated. At the same time, data quality assessment and improvement is more challenging. The following types of information systems can be distinguished based on their degree of integration:

—In a *monolithic information system*, applications are single-tier and do not provide data access services. Although data are usually stored in a database that can be queried, separate applications do not share data. This can cause data duplication, possibly affecting all quality dimensions.

—A *data warehouse (DW)* is a centralized collection of data retrieved from multiple databases. Data warehouses are periodically refreshed with updated data from the original databases by procedures automatically extracting and aligning data. Data are physically integrated, since they are reformatted according to the data warehouse schema, merged, and finally stored, in the data warehouse.

—A *distributed information system* is a collection of application modules coordinated by a workflow. Applications are typically divided in tiers, such as presentation, application logic, and data management, and export data access functionalities at different tiers. Data can be stored in different databases, but interoperability is guaranteed by the logical integration of their schemas.

—A *cooperative information system* (CIS) can be defined as a large-scale information system that interconnects multiple systems of different and autonomous organizations sharing common objectives [De Michelis et al. 1997]. Cooperation with other information systems requires the ability to exchange information. In CISs, data are not logically integrated, since they are stored in separate databases according to different schemas. However, applications incorporate data transformation and exchange procedures that allow interoperability and cooperation among common processes. In other words, integration is realized at a process level.

—In the literature, the term *Web Information System* (WIS) [Isakowitz et al. 1998] is used to indicate any type of information adopting Web technologies. From a technical perspective a WIS is a client/server application. Such systems typically use structured, semi structured, and unstructured data, and are supported by development and management tools based on techniques specific to each type of data.

—In a *peer-to-peer information system* (P2P), there is no distinction between clients and servers. The system is constituted by a set of identical nodes that share data and application services in order to satisfy given user requirements collectively. P2P systems are characterized by a number of properties: no central coordination, no central database, no peer has a global view of the system, Peers are autonomous and can dynamically connect or disconnect from the system. However, peers typically share common management procedures.

3. COMPARISON OF METHODOLOGIES

This section compares methodologies based on the classification criteria discussed in the previous section. Table III shows the list of methodologies considered in this paper identified by acronyms together with the extended name of the methodology and the

Table III. Methodologies Considered in the Article

Methodology Acronym	Extended Name	Main Reference
TDQM	Total Data Quality Management	Wang 1998
DWQ	The Datawarehouse Quality Methodology	Jeusfeld et al. 1998
TIQM	Total Information Quality Management	English 1999
AIMQ	A methodology for information quality assessment	Lee et al. 2002
CIHI	Canadian Institute for Health Information methodology	Long and Seko 2005
DQA	Data Quality Assessment	Pipino et al. 2002
IQM	Information Quality Measurement	Eppler and Münzenmaier 2002
ISTAT	ISTAT methodology	Falorsi et al 2003
AMEQ	Activity-based Measuring and Evaluating of product information Quality (AMEQ) methodology	Su and Jin 2004
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality	Loshin 2004
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco et al. 2004
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis and Batini 2004
CDQ	Comprehensive methodology for Data Quality management	Batini and Scannapieco 2006

Table IV. Methodologies and Assessment Steps

Step/Meth Acronym	Data Analysis	DQ Requirement Analysis	Identification of Critical Areas	Process Modeling	Measurement of Quality	Extensible to Other Dimensions and Metrics
TDQM	+		+	+	+	Fixed
DWQ	+	+	+		+	Open
TIQM	+	+	+	+	+	Fixed
AIMQ	+		+		+	Fixed
CIHI	+		+			Fixed
DQA	+		+		+	Open
IQM	+				+	Open
ISTAT	+				+	Fixed
AMEQ	+		+	+	+	Open
COLDQ	+	+	+	+	+	Fixed
DaQuinCIS	+		+	+	+	Open
QAFD	+	+	+		+	Fixed
CDQ	+	+	+	+	+	Open

main reference. The acronym will be used to identify each methodology in the remainder of this article.

Costs, dimensions, and phases represent the most discriminating criteria, leading to the identification of four types of methodologies, which are discussed in Section 3.7, and highlighting the fundamental differences among them. In the rest of the section, methodologies are compared in depth along the perspectives identified in Section 2.

3.1. Methodologies, Phases, and Steps

Tables IV, V, and VI show the phases and steps addressed by each methodology. A methodology has been considered to include a phase or a step if it provides at least a discussion of the corresponding phase or step, and possibly, methodological guidelines and original techniques. For example, DWQ generically refers to a *process modeling* step, but does not provide execution details. For the same reason, the *measurement of*

Table V. Methodologies and Improvement Steps-Part 1

Step/Meth. Acronym	Evaluation of Costs	Assignment of Process Responsibilities	Assignment of Data Responsibilities	Selection Strategies and Techniques	Identification the Causes of Errors
TDQM	+	+	+	+	+
DWQ	+		+	+	+
TIQM	+	+	+	+	+
DQA					+
ISTAT				+	+
AMEQ					+
COLDQ	+			+	+
DaQuinCIS				+	+
CDQ	+	+	+	+	+

Table VI. Methodologies and Improvement Steps-Part 2

Step/Meth. Acronym	Process Control	Design of data Improvement Solutions	Process Redesign	Improvement Management	Improvement Monitoring
TDQM			+	+	+
DWQ		+		+	
TIQM		+	+		+
DQA					
ISTAT		+	+		
AMEQ					+
COLDQ	+	+	+		+
DaQuinCIS					
CDQ	+	+	+		

quality step is not associated with CIHI in Table IV. Assessment and improvement are discussed separately in the next two sections.

3.1.1. The Assessment Phase. Table IV compares the steps followed by different methodologies in the assessment phase. In general, methodologies refer to the steps classified in Section 2.1, although with different names. However, it is not difficult to recognize a name correspondence by analyzing the objectives of the step. For example, CIHI discusses a methodological step that identifies the databases with a quality level below a given acceptability threshold. This step has a clear correspondence with the *find critical areas* step of Section 2.1.

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. Different measurement approaches meet the specific requirements of different organizational contexts, processes, users or services. Only a few methodologies consider the *DQ requirements analysis* step, identifying DQ issues and collecting new target quality levels from users. This step is particularly relevant for evaluating and solving conflicts in target DQ levels from different stakeholders. For example, QAFD recommends the collection of target quality levels from different types of experts, including business experts and financial operators, but does not help the reconciliation of incompatible DQ levels. A few methodologies support *process modeling*. Note that with the exception of AMEQ, the methodologies supporting process modeling also adopt a process-driven strategy for the improvement phase (see next section).

Table VII. Methodologies and Types of Strategies

Strategy/Meth. Acronym	Data-driven	Process-driven
TDQM		Process Redesign
DWQ	Data and schema integration	
TIQM	Data cleansing Normalization Error localization and correction	Process Redesign
ISTAT	Normalization Record linkage	Process Redesign
COLDQ	Cost optimization	Process Control Process Redesign
DaQuinCIS	Source trustworthiness Record Linkage	
CDQ	Normalization Record Linkage Data and schema integration Error localization and correction	Process Control Process Redesign

The last column of Table IV specifies whether the methodology allows extensibility to dimensions (and metrics) other than those explicitly dealt with in the methodology. For example, CDQ explicitly mentions dimensions among those that will be described in Section 3.3, but the approach can be easily generalized to other dimensions. On the contrary, ISTAT provides detailed measurement and improvement procedures for accuracy, completeness, and consistency, and consequently, the whole approach is strictly hardwired to such dimensions.

Note that the methodologies that address both the *process modeling* and *measurement of quality* steps, are based on the *fitness for use* approach. They evaluate the quality of data along the processes in which they are used and, thus mainly provide subjective measures.

3.1.2. The Improvement Phase. Tables V and VI compare the improvement steps of different methodologies.

The *identification of the causes of errors* is the most widely addressed improvement step. DQA emphasizes the importance of the *identification of the causes of errors* step, but it does not discuss its execution. Similarly, DWQ refers to a mathematical model based on the concept of dependency to support the *identification of the causes of errors* step, but the definition of the model is presented as ongoing work and is not provided.

Only six methodologies address multiple improvement steps, as confirmed by Table VII. Improvement activities are mostly based on *process redesign*, with the exception of the DWQ methodology, which provides an extension of the Goal Question Metric [Basili et al. 1994] initially proposed in the software engineering field. The *cost evaluation* step is usually mandatory in DQ methodologies. This step is considered critical for measuring the economic advantage of improvement solutions and to choose the most efficient improvement techniques. In contrast, the *management of the improvement solution* step is explicitly performed only by TDQM. Other methodologies refer to the broad range of management techniques and best practices available from the change management field [Kettinger and Grover 1995]. Furthermore, it is possible to repeat the assessment phase of the methodology in order to evaluate the results of the improvement phase. As an example, DQA explicitly recommends the application of previous methodological steps to evaluate the effectiveness of improvement.

Finally, the relationship among data quality, process, and organization is considered by TIQM, TDQM, and CDQ. These methodologies thoroughly discuss the *assignment* *of responsibilities on processes* and *data*. These steps are supported by the results of the state reconstruction phase. CDQ discusses a set of matrices to represent the relationship among processes, organizational units, and databases, which are produced during the state reconstruction phase and are subsequently used in the *assignment of responsibilities* steps.

3.2. Methodologies, Strategies, and Techniques

Table VII shows the strategies and techniques adopted by different methodologies. A methodology is associated with a strategy if it provides guidelines to select and design corresponding techniques.

Notice that the column labelled *Process-driven* in Table VII provides the same information as columns, *Process control* and *Process redesign* of Table VI. The column labelled *Data-driven* explicitly mentions the data-driven techniques implicitly considered in Tables V and VI.

Table VII shows that five DQ methodologies adopt mixed strategies, variously combining data-driven and process-driven techniques. The methodology applying the wider range of data- and process-driven techniques is TIQM. Conversely, TDQM provides guidelines to apply process-driven strategies by using the Information Manufacturing Analysis Matrix [Ballou et al. 1998], which suggests when and how to improve data.

It is worth noting that a methodology exclusively adopting either a data- (as for DWQ and DaQuinCIS) or a process-driven strategy, may not be flexible for organizations that have DQ practices. The only methodology that explicitely addresses this issue is CDQ, which jointly selects data- and process-driven techniques. The selection of the most suitable strategy and technique is based on domain-dependent decision variables [Batini et al. 2008].

Normalization, record linkage, data and schema integration, represent the data-driven techniques most widely adopted in DQ methodologies, while process redesign, as discussed in previous section, is most relevant in process-driven methodologies. We now discuss specific contributions related to the data- and process-driven techniques considered in Section 2.2.

3.2.1. Data-Driven Techniques.

Normalization techniques have been proposed in several domains, including census and territorial data domains. Both ISTAT and CDQ provide normalization techniques improving DQ by comparing data with look-up tables and defining a common metaschema. For example, the ISTAT methodology uses the national street registry as a lookup table for territorial data.

Record linkage has been investigated in the database research since the '50s and has been applied in many contexts such as healthcare, administrative, and census applications. In such contexts, it is crucial to produce efficient computer-assisted matching procedures that can reduce the use of clerical resources, and at the same time, minimize matching errors. CDQ discusses three types of record linkage techniques:

(1) *Probabilistic techniques*, based on the broad set of methods developed over the past two centuries within statistics and probability theory, ranging from Bayesian networks to data mining.
(2) *Empirical techniques* that make use of algorithmic techniques such as sorting, tree analysis, neighbor comparison, and pruning.
(3) *Knowledge-based techniques*, extracting knowledge from files and applying reasoning strategies.

Criteria for choosing among these three types of techniques are discussed within the CDQ methodology. The DaQuinCIS project has developed a specific record linkage technique [Bertolazzi et al. 2003]. In the DaQuinCIS platform, record linkage is performed in two phases: (1) first, record linkage aligns different copies of the same entities in different data sources; (2) second, record linkage also supports the query processing phase by identifying the same instances in the query results returned by each data source. The record linkage method is based on the Sorted Neighborhood method [Hernandez and Stolfo 1998], but some new features are introduced:

—the matching key is automatically selected instead of being selected by the key designer;

—the matching algorithm is based on a function that normalizes a classic edit distance function upon string lengths.

Data and schema integration [Lenzerini 2002] is a broad area of research that partially overlaps with data quality. Data-driven improvement techniques applied in the methodologies are often based on the use of new data to improve the quality of a given data collection. As a consequence, DQ improvement techniques focus primarily on instance-level heterogeneities, in order to identify similar records, detect conflicting values, and select a single final instance.

Methodologies that address instance-level heterogeneities are DWQ, ISTAT, DaQuinCIS, and CDQ. In DWQ, heterogeneous information sources are first made accessible in a uniform way through extraction mechanisms called *wrappers*, then *mediators* take on the task of information integration and conflict resolution. The resulting standardized and integrated data are stored as materialized views in the data warehouse.

ISTAT suggests how to resolve heterogeneities among data managed by different public agencies by adopting a common model for representing the format of exchanged data, based on the XML markup language. In this way, the comprehension of heterogeneities among agencies is made easier, while the solution of such heterogeneities is left to bilateral or multilateral agreements.

In DaQuinCIS, instance-level heterogeneities among different data sources are dealt with by the DQ broker. Different copies of the same data received as responses to the request are reconciled by the DQ broker, and a best-quality value is selected.

CDQ follows an approach similar to ISTAT, with more emphasis on the autonomy of organizations in the cooperative system. In fact, the resolution of heterogeneities in the case studies, proposed as best practices, is performed through record linkage on a very thin layer of data, namely the identifiers. All other data are reconcilied only in case of autonomous decisions of the agencies involved.

3.2.2. Process-Driven Techniques. Methodologies addressing the process redesign step tend to borrow corresponding techniques from the literature on business process reengineering (BPR) [Muthu et al. 1999; Hammer 1990]. TDQM represents an exception in this respect, as it proposes an original process redesign control approach that is referred to as an “information manufacturing system for the Information Product” [Ballou et al. 1998]. This methodology proposes the Information Production Map (IP-MAP) model [Shankaranarayan et al. 2000] that is used to model the information products managed by the manufacturing processes. An information production map is a graphical model designed to help analysts to visualize the information production process, identify the ownership of process phases, understand information and organizational boundaries, and estimate the time and quality metrics associated with the current production process. The description of processes is a mandatory activity, consistent with the general orientation of process-driven strategies. After modelling and assessing the information production process, new process control activities are identified and/or process redesign decisions are taken.

Complex solutions such as IP-MAP cannot always be adopted due to their high costs and, in some cases, the practical unfeasibility of a thorough process modeling step. For this reason, other methodologies adopt less formal, but more feasible solutions. For example, CDQ is based on a set of matrices that describe the main relationships among data, information flows, processes, and organizational units. The relationship between organizational units and processes has also been modeled in extensions of IP-MAP proposed in the literature [Scannapieco et al. 2002].

3.3. Methodologies and Dimensions

Table VIII shows the quality dimensions considered by the methodologies surveyed in this article. In Table VIII, a dimension is associated with a methodology, if the methodology provides a corresponding definition. For each methodology's dimensions, we address the corresponding references (see Table III).

Notice the large variety of dimensions defined in the methodologies, which confirms the complexity of the *data quality* concept. This is not surprising, since nowadays a large number of phenomena can be described in terms of data. Multiple classifications of quality dimensions are proposed by the methodologies. TIQM classifies dimensions as *inherent* and *pragmatic*. COLDQ distinguishes among *schema*, *data*, *presentation*, and *information policy* dimensions. CIHI provides a two-level classification in terms of dimensions and related *characteristics*. CDQ proposes *schema* and *data* dimensions.

Table IX shows the metrics provided for quality dimensions by different methodologies. We do not include metrics for semantic accuracy because the two methodologies addressing it, namely QAFD and CDQ, do not provide specific measurement methods. In general, multiple metrics are defined for each dimension, and each dimension accordingly has multiple entries in the table. Note that subjective metrics such as user surveys have been defined for almost all quality dimensions. Different metrics for the same dimension are identified by acronyms, which are used in Table X to associate them with the methodologies in which they are used and/or defined.

The last column of Table X provides for each dimension and each metric associated with the dimension, (1) the number of methodologies that use the metrics, and (2) the total number of methodologies that mention the corresponding dimension. The ratio between these values measures the degree of consensus on dimension metrics among methodologies. Such consensus is high for accuracy, completeness, and consistency, while it is significantly lower for two of the time-related dimensions, timeliness and currency, and almost all other dimensions.

The majority of metrics with only one occurrence in methodologies are mentioned in IQM, which analyzes the quality of Web information. Such metrics are defined by considering the measurement tools that are available in the specific Web context. For example, using a site analyzer, it is possible to assess dimensions such as accessibility, consistency, timeliness, conciseness, and maintainability. Traffic analyzers can be used to assess applicability and convenience, while port scanners are useful to assess security. The high number of measurement tools in the Web context results in a high number of metrics specific to IQM.

AIMQ has several specific metrics and dimensions. This is due to the top-down approach adopted in AIMQ in the definition of dimensions and metrics, which uses two different classifications (not represented in Table VIII): (1) product vs. service quality, and (2) conforms to specification vs. meets or exceeds customer expectations, leading to a widely scattered set of related dimensions/metrics.

Note that the AIMQ methodology uses only subjective metrics to assess quality dimensions. In AIMQ, data quality is mainly assessed by means of questionnaires that

Table VIII. Methodologies and Quality Dimensions

Acronym	Data Quality Dimension
TDQM	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of manipulation, Value added, Free of error, Interpretability, Objectivity, Relevance, Reputation, Security, Timeliness, Understandability
DWQ	Correctness, Completeness, Minimality, Traceability, Interpretability, Metadata Evolution, Accessibility (System, Transactional, Security), Usefulness (Interpretability), Timeliness (Currency, Volatility), Responsiveness, Completeness, Credibility, Accuracy, Consistency, Interpretability
TIQM	Inherent dimensions: Definition conformance (consistency), Completeness, Business rules conformance, Accuracy (to surrogate source), Accuracy (to reality), Precision, Nonduplication, Equivalence of redundant data, Concurrency of redundant data, Pragmatic dimensions: accessibility, timeliness, contextual clarity, Derivation integrity, Usability, Rightness (fact completeness), cost.
AIMQ	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of operation, Freedom from errors, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability
CIHI	Dimensions: Accuracy, Timeliness Comparability, Usability, Relevance Characteristics: Over-coverage, Under-coverage, Simple/correlated response variance, Reliability, Collection and capture, Unit/Item non response, Edit and imputation, Processing, Estimation, Timeliness, Comprehensiveness, Integration, Standardization, Equivalence, Linkage ability, Product/Historical comparability, Accessibility, Documentation, Interpretability, Adaptability, Value.
DQA	Accessibility, Appropriate amount of data, Believability, Completeness, Freedom from errors, Consistency, Concise Representation, Relevance, Ease of manipulation, Interpretability, Objectivity, Reputation, Security, Timeliness, Understandability, Value added.
IQM	Accessibility, Consistency, Timeliness, Conciseness, Maintainability, Currency, Applicability, Convenience, Speed, Comprehensiveness, Clarity, Accuracy, Traceability, Security, Correctness, Interactivity.
ISTAT	Accuracy, Completeness, Consistency
AMEQ	Consistent representation, Interpretability, Case of understanding, Concise representation, Timeliness, Completeness Value added, Relevance, Appropriateness, Meaningfulness, Lack of confusion, Arrangement, Readable, Reasonability, Precision, Reliability, Freedom from bias, Data Deficiency, Design Deficiency, Operation, Deficiencies, Accuracy, Cost, Objectivity, Believability, Reputation, Accessibility, Correctness, Unambiguity, Consistency
COLDQ	Schema: Clarity of definition, Comprehensiveness, Flexibility, Robustness, Essentialness, Attribute granularity, Precision of domains, Homogeneity, Identifiability, Obtainability, Relevance, Simplicity/Complexity, Semantic consistency, Syntactic consistency. Data: Accuracy, Null Values, Completeness, Consistency, Currency, Timeliness, Agreement of Usage, Stewardship, Ubiquity, Presentation: Appropriateness, Correct Interpretation, Flexibility, Format precision, Portability, Consistency, Use of storage, Information policy: Accessibility, Metadata, Privacy, Security, Redundancy, Cost.
DaQuinCIS	Accuracy, Completeness, Consistency, Currency, Trustworthiness
QAFD	Syntactic/Semantic accuracy, Internal/External consistency, Completeness, Currency, Uniqueness.
CDQ	Schema: Correctness with respect to the model, Correctness with respect to Requirements, Completeness, Pertinence, Readability, Normalization, Data: Syntactic/Semantic Accuracy, Semantic Accuracy, Completeness, Consistency, Currency, Timeliness, Volatility, Completability, Reputation, Accessibility, Cost.

include 4–6 independent items for each quality dimension. Items have the following general structure: "This information is (attribute or phrase)." For example, completeness is associated with six items, including: (1) this information provides all necessary values, (2) this information is sufficiently complete for our needs, (3) this information fulfils the needs of our tasks.

Table IX. Dimensions and Metrics

Dimensions	Name	Metrics Definition
Accuracy	Acc1	Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one Syntactic Accuracy=Number of correct values/number of total values
	Acc2	Number of delivered accurate tuples
	Acc3	User Survey - Questionnaire
Completeness	Compl1	Completeness = Number of not null values/total number of values
	Compl2	Completeness = Number of tuples delivered/Expected number
	Compl3	Completeness of Web data = (T$_{max}$- T$_{current}$)* (Completeness$_{Max}$- Completeness$_{Current}$)/2
	Compl4	User Survey - Questionnaire
Consistency	Cons1	Consistency = Number of consistent values/number of total values
	Cons2	Number of tuples violating constraints, number of coding differences
	Cons3	Number of pages with style guide deviation
	Cons4	User Survey - Questionnaire
Timeliness	Time1	Timeliness = (max (0; 1-Currency/Volatility))^{s}
	Time2	Percentage of process executions able to be performed within the required time frame
	Time3	User Survey - Questionnaire
Currency	Curr1	Currency = Time in which data are stored in the system - time in which data are updated in the real world
	Curr2	Time of last update
	Curr3	Currency = Request time- last update
	Curr4	Currency = Age + (Delivery time- Input time)
	Curr5	User Survey - Questionnaire
Volatility	Vol1	Time length for which data remain valid
Uniqueness	Uni1	Number of duplicates
Appropriate amount of data	Appr1	Appropriate Amount of data = Min ((Number of data units provided/Number of data units needed); (Number of data units needed/Number of data units provided))
	Appr2	User Survey - Questionnaire
Accessibility	Access1	Accessibility = max (0; 1-(Delivery time - Request time)/(Deadline time - Request time))
	Access2	Number of broken links - Number of broken anchors
	Access3	User Survey - Questionnaire
Credibility	Cred1	Number of tuples with default values
	Cred2	User Survey - Questionnaire
Interpretability	Inter1	Number of tuples with interpretable data, documentation for key values
	Inter2	User Survey - Questionnaire
Usability	Usa1	User Survey - Questionnaire
Derivation Integrity	Integr1	Percentage of correct calculations of derived data according to the derivation formula or calculation definition
Conciseness	Conc1	Number of deep (highly hierarchic) pages
	Conc2	User Survey - Questionnaire
Maintainability	Main1	Number of pages with missing meta-information
Applicability	App1	Number of orphaned pages
	App2	User Survey - Questionnaire
Convenience	Conv1	Difficult navigation paths: number of lost/interrupted navigation trails
Speed	Speed1	Server and network response time
Comprehensiveness	Comp1	User Survey - Questionnaire
Clarity	Clar1	User Survey - Questionnaire
Traceability	Trac1	Number of pages without author or source
Security	Sec1	Number of weak log-ins
	Sec2	User Survey - Questionnaire
Correctness	Corr1	User Survey - Questionnaire
Objectivity	Obj1	User Survey - Questionnaire
Relevancy	Rel1	User Survey - Questionnaire
Reputation	Rep1	User Survey - Questionnaire
Ease of operation	Ease1	User Survey - Questionnaire
Interactivity	Interact1	Number of forms - Number of personalizable pages

Table X. Methodologies and Quality Metrics

	TDQM	DWQ	TIQM	AIMQ	CIHI	DQA	IQM	ISTAT	AMEQ	COLDQ	DaQuinCIS	QAFD	CDQ	#metr/#dim
Acc1	X		X			X		X	X	X	X	X	X	9/13
Acc2		X												1/13
Acc3				X	X		X							2/13
Compl1	X			X		X			X	X	X	x	x	7/12
Compl2		X	X											2/12
Compl3													X	1/12
Compl4			X											2/12
Cons1	X					X								6/10
Cons2		X												1/10
Cons3							X							1/10
Cons4				X	X									2/10
Time1						X				X			X	3/7
Time2			X							X				2/7
Time3				X	X									2/7
Curr1								X			X			2/8
Curr2	X						X					X		2/8
Curr3				X	X									1/8
Curr4													X	1/8
Curr5				X	X									2/8
Vol1	X												X	2/2
Uni1			X									X		1/2
Appr1						X								1/2
Appr2				X										1/2
Access1							X							1/4
Access2								X						1/4
Access3			X	X										2/4
Cred1		X												1/2
Cred2				X										1/2
Inter1		X												1/2
Inter2				X										1/2
Usa1			X											1/1
Integr1			X											1/1
Conc1														1/2
Conc2							X							1/ 2
Main1				X										1/1
App1														1/1
App2							X							1/1
Conv1							X							1/1
Speed1							X							1/1
Comp1				X			X			X				3/3
Clar1			X				X			X				3/3
Trac1							X							1/1
Sec1							X							1/1
Sec2				X										1/1
Corr1							X							1/1
Obj1				X										1/1
Rel1				X										1/1
Rep1				X										1/1
Ease1				X										1/1
Interact1							X							1/1

Finally, the metrics provided by the DWQ methodology have a tight relation with the processes using data. All dimensions are evaluated along the expectations of the users of a particular process. For example, completeness is defined as the ratio of not null values to the number of values required by a specific process, as opposed to the total number of values stored in a database (see Section 2.3).

Table XI. Comparison Between English and Loshin Classifications

Methodologies/ Cost Categories	TIQM	COLDQ
Cost of assessment DQ activities	Assessment or inspection costs: • Information quality analysis software costs • People time in the assessment process	Detection costs
Cost of improvement DQ activities	Process improvement and defect prevention cost	Data correction costs Data maintenance costs: • Acquisition overhead costs • Decay costs • Infrastructure costs Process improvement costs
Process costs of poor data quality	Process failure costs: • Unrecoverable costs • Liability and exposure costs • Recovery costs of unhappy customers Information scrap and rework: • Redundant data handling and support costs • Costs of hunting or chasing missing information • Business rework costs • Workaround costs and decreased productivity • Data verification costs • Software rewrite costs • Data cleaning and correction costs • Data cleaning software costs	Operational impacts: • Rollback costs • Rework costs • Prevention costs • Warranty costs Tactical and strategic impacts: • Delay costs • Preemption costs • Idling costs • Increased difficulty costs • Lost difficulty costs • Organizational mistrust costs • Misalignment costs
Opportunity costs of poor data quality	Lost and missed opportunity costs: • Lost opportunity costs • Missed opportunity costs • Lost shareholder value	Lost revenue costs: • Spin costs • Reduction costs • Attrition costs • Blockading costs

3.4. Methodologies and Costs

The cost dimension is considered only in TIQM, COLDQ, and CDQ. In this section we analyze costs from two different points of view: (1) cost classifications, and (2) criteria provided for cost quantification.

3.4.1. Cost Classifications. Both TIQM [English 1999] and COLDQ [Loshin 2004] provide detailed classifications for costs. A third classification is provided by Eppler and Helfert [2004].

Table XI compares the TIQM and COLDQ classifications. In TIQM, data quality costs correspond to the costs of business processes and data management processes due to poor data quality. Costs for information quality assessment or inspection measure data quality dimensions to verify that processes are performing properly. Finally, process improvement and defect prevention costs involve activities to improve the quality of data, with the goal of eliminating, or reducing, the costs of poor data quality. Costs due to poor data quality are analyzed in depth in the TIQM approach, and are subdivided into three categories:

(1) *Process failure costs* are incurred when poor quality data causes a process not to perform properly. As an example, inaccurate mailing addresses cause correspondence to be misdelivered.

(2) *Information scrap and rework*. When data is of poor quality, they involve several types of defect management activities, such as reworking, cleaning, or rejecting.

(3) *Loss and missed opportunity costs* correspond to the revenues and profits lost because of poor data quality. For example, due to low accuracy of customer e-mail addresses, a percentage of customers already acquired cannot be reached by periodic advertising campaigns, resulting in lower revenues, roughly proportional to the decrease of the accuracy of addresses.

COLDQ analyzes the costs of low data quality, classifying them according to their domain impact, namely:

—the *operational domain*, which includes the components of the information processing system and the operating costs of the system;
—the *tactical domain*, which attempts to address and solve problems before they arise;
—the *strategic domain*, which stresses long-term decisions.

For both the operational and tactical/strategic impact several cost categories are introduced. Here, we describe the operational impact costs:

—*rollback costs* are incurred when work that has been performed needs to be undone;
—*rework costs* are incurred when a processing stage must be repeated;
—*prevention costs* arise when a new activity is implemented to take the actions necessary to prevent operational failure due to a data quality problem;
—*warranty costs* are related to guarantees against losses.

Finally, we mention that CDQ proposes a classification that reconciles the heterogeneities among TIQM, COLDQ, and Eppler and Helfert [2004]. For details, see Batini and Scannapieco [2006].

3.4.2. Criteria for Cost Quantification. The assessment of the total cost of data quality supports the selection of the types of data quality activities to be performed (see Section 2.2) and their prioritization. TIQM, COLDQ, and CDQ are the only methodologies providing criteria for this activity. In TIQM, selection and prioritization are achieved with the following steps:

—identify current users and uses of data;
—list the errors that negatively affect data quality;
—identify the business units most often impacted by poor quality data;
—estimate the direct cost of the current data quality program;
—estimate the costs of data errors for all users and uses of data, grouped by business unit;
—use costs to justify and prioritize data quality initiatives, including the institutionalization of a continuous quality improvement program.

Each type of error occurs with a given frequency and involves a cost. Note that the cost of different error categories is a contingent value that varies with the process that makes use of the data. Models for process representation allow the identification of the activities affected by data errors. Since activities are typically associated with a total organizational cost, the cost of rework can provide quantitative and objective estimates.

COLDQ focuses on the evaluation of the cost associated with poor data quality, as an argument for supporting the investment in a knowledge management program. The evaluation is achieved with the following steps:

—map the information chain to understand how information flows within the organization;

Table XII. Methodologies and Types of Data

Type of data/Meth. acronym	Structured	Semistructured
TDQM	x	x
DWQ	x	
TIQM	x	Implicitly considered
AIMQ	x	Implicitly considered
CIHI	x	x
DQA	x	
IQM	x	x
ISTAT	x	x
AMEQ	x	Implicitly considered
COLDQ	x	Implicitly considered
DaQuinCIS	x	x
QAFD	x	
CDQ	x	x

—identify current users and uses of data;
—identify the impact of the quality of data on customers;
—isolate flawed data by locating critical areas;
—identify the impact domain associated with each instance of poor data quality;
—characterize the economic impact based on the ultimate effects of bad data;
—aggregate the totals to evaluate the overall economic impact;
—identify opportunities for improvement.

The result is called the data quality scorecard. It summarizes the cost associated with poor data quality and can be used as a tool to find the best solutions for improvement.

In CDQ the minimization of the cost of the data quality program is the main criterion for choosing among alternative improvement processes. First, different improvement processes are identified as paths of data- and process-driven techniques applied to the data bases, data flows, and document bases involved in the improvement. Then, the costs of the different processes are evaluated and compared, and the minimum-cost process is selected.

3.5. Methodologies and Types of Data

We observed in Section 2.5 that the types of data influence the DQ dimensions and the assessment and improvement techniques. Table XII associates the types of data classified in Section 2.5 and DQ methodologies. Most methodologies address structured data, while only a few also address semistructured data. In Table XII we have imputed *implicitely considered* when the methodology does not explicitely mention the type of data, but phases and steps can be applied to it. For example, AIMQ uses the generic term *information*, and performs qualitative evaluation through questions that apply to structured data, but may refer to any type of data, including unstructured data.

Concerning *semistructured data*, ISTAT considers the standardization of address data formats and their expression in a common XML schema. This schema is implemented to minimize discrepancies across agencies and allow interoperability. In the DaQuinCIS methodology [Scannapieco et al. 2004], a model associating quality values with XML documents is proposed. The model, called *Data and Data Quality* (D^2Q), is intended to be used in the context of data flows exchanged by different organizations in a cooperative information system. In the exchange of data, the quality of data flows becomes critical to avoid error propagation. D^2Q can be used in order to certify

Table XIII. Methodologies and Information Systems

Type of inf.syst. /Meth. acronym	Monolithic	Distributed	DataWarehouse	Cooperative	Web
TDQM	focused	implicitly considered			
DWQ			strongly focused		
TIQM	focused	focused			
AIMQ	focused	implicitly considered			
CIHI	focused	focused			
DQA	focused	implicitly considered			
IQM					strongly focused
ISTAT	focused	focused		strongly focused	
AMEQ	focused				
COLDQ	focused				
DaQuinCIS	focused	focused		strongly focused	
QAFD	focused				
CDQ	focused	focused		strongly focused	

the accuracy, completeness, currency, and consistency of data. The model is semistructured, thus allowing each organization to export the quality of its data with a certain degree of flexibility. The quality values can be associated with various elements of the data model, ranging from individual data values to the whole data source.

CDQ attempts an extension to semistructured data of steps and techniques originally developed for structured data. For example, all data types, both structured and semistructured, are surveyed in the state reconstruction phase. A specific data-driven improvement technique is also proposed for unstructured data, called *data profiling*. This technique is used to relate a text file to a database schema by discovering recurring patterns inside text [Aiken 1996].

3.6. Methodologies and Types of Information Systems

Table XIII shows to what extent the different methodologies deal with the types of information systems introduced in Section 2.6. The table adopts a four-value scale, where: (1) *strongly focused* means that the whole organization of the methodology is conceived and tailored to the corresponding type of information system, while providing generic guidelines for other types of information systems, (2) *focused* means that the methodology provides detailed guidelines and techniques for the corresponding type of information system, (3) *implicitly considered* has the same meaning as in Table XII, and (4) a *missing value* indicates that the methodology either provides generic guidelines, or does not explicitly address the corresponding type of information system. Web information systems are included in the table, since one methodology is strongly focused on them; further issues related to such systems will be discussed in Section 4. No methodology mentions P2P systems. They will be considered in the conclusions and open issues (Section 4).

It can be observed that AMEQ, COLDQ, and QAFD focus on monolithic information systems. They typically consider structured data sets within a single system, and ignore DQ issues raised by data exchanges among separate applications or organizations.

TIQM can be applied to both monolithic and distributed systems, because when the data architecture is analyzed, both a centralized and a distributed system are provided as case studies. CIHI is considered focused on distributed systems since national databases are mentioned in the description of the methodology.

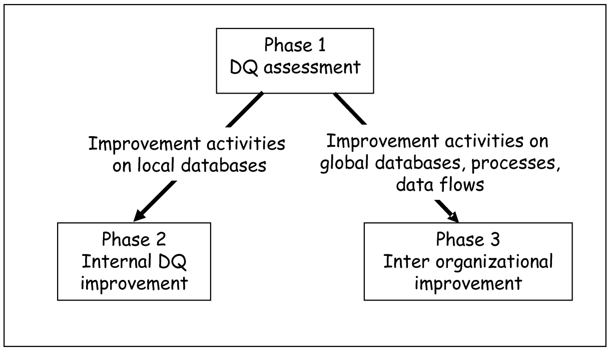

Phase 1
DQ assessment
Improvement activities on local databases
Improvement activities on global databases, processes, data flows
Phase 2
Internal DQ improvement
Phase 3
Inter organizational improvement

Fig. 3. General view of the Istat methodology.

Other methodologies, ISTAT, DaQuinCIS, and CDQ, can be applied to multiple types of systems. Each methodology provides detailed guidelines for the most complex type of system, the cooperative information system, and can therefore also be applied to monolithic and distributed systems. The methodologies follow different approaches for dealing with multiple data flows and heterogeneous databases.

ISTAT has a strong interorganizational approach, as it has been conceived for the Italian public administration, which is characterized by a distributed structure with highly autonomous administrations (this is typical of many other countries). The three main phases of the methodology are shown in Figure 3.

The assessment phase (Phase 1) identifies the most relevant activities to be performed in the improvement phase (Phases 2 and 3):

(1) Phase 2 acts on local databases owned and managed by different administrations. Tools are distributed to perform DQ activities autonomously, and courses are offered to improve local DQ skills.

(2) Phase 3 concerns the overall cooperative information system of a set of administrations, in terms of exchanged data flows and central databases set up for coordination purposes. These activities are centrally planned and coordinated.

DaQuinCIS provides a framework that offers several services and tools for cooperative information systems (see Figure 4). The *data quality broker*, described in Section 3.2, is the core of the architecture, and is responsible for the reconciliation of heterogeneous responses to queries. The *quality notification service* is a publish/subscribe engine used as a general message bus between the architectural components of different cooperating organizations [Scannapieco et al. 2004]. It allows quality-based subscriptions for organizations to be notified on changes of the quality of data. The best-quality value selected by the broker, mentioned in Section 3.2, is proposed to requesting organizations, which can choose to replace their data with higher quality data (*quality improvement function*). The *quality factory* component (see Figure 4) is responsible for evaluating the quality of the internal data of each organization (the interested reader can refer to Cappiello et al. [2003b]). Requests from external users or information systems are processed in the quality factory by the *quality analyzer*, which performs a static analysis of the values of the data quality dimensions associated with requested data, and compares them with benchmark quality parameters contained in the *quality repository*. The *rating service* will be addressed in Section 4.

Other contributions in the literature that provide tools for quality-based query processing and instance-level conflict resolution for cooperative information systems are

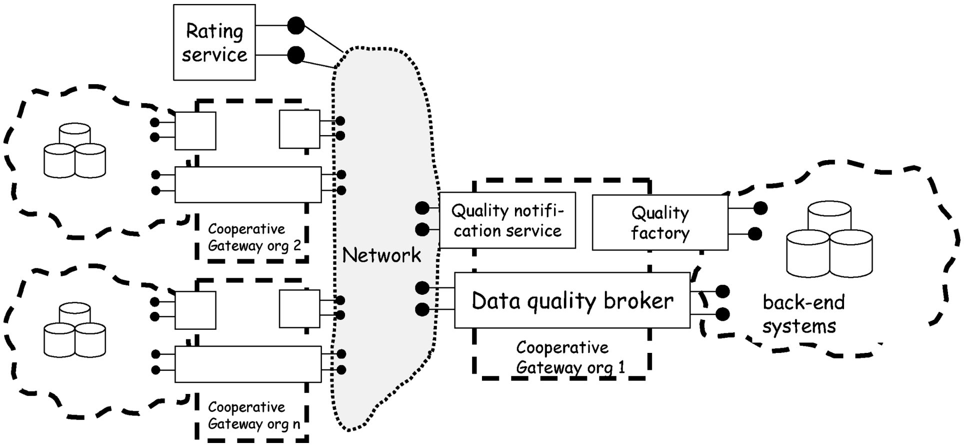

Rating service
Cooperative Gateway org 2
Network
Quality notifi-cation service
Quality factory
Data quality broker
back-end systems
Cooperative Gateway org 1
Cooperative Gateway org n

Fig. 4. The architecture of the DaQuinCIS framework.

Fusionplex [Motro and Anokhin 2005], iFuice [Rahm et al. 2005], and HumMer [Bilke et al. 2005].

We now comment on the methodologies that are strongly focused on specific types of information systems, namely DWQ and IQM. DWQ is specifically oriented to data warehouses. In Jarke et al. [1995], it is observed that most researches have studied DW in their role as buffers of materialized views, mediating between update-intensive OLTP systems and query-intensive decision support. This neglects the organizational role of data warehousing as a means of obtaining a centralized control of information flows. As a consequence, a large number of quality issues relevant for DW cannot be expressed with traditional DW meta models. DWQ makes two contributions towards solving these problems. First, the metadata about DW architectures are enriched with explicit enterprise models. Second, mathematical techniques for measuring or optimizing several aspects of DW quality are proposed.

The Artkos tool also contributes to address the DQ issues of data warehouses [Vassiliadis et al. 2001]. Artkos proposes a metamodel made of several entities, among them: (1) *activities*, atomic units of data processing work; (2) *scenarios*, sets of activities to be executed together; and (3) *quality factors*, defined for each activity, corresponding to the dimensions and metrics described in Section 2.3.

IQM is strongly focused on Web information systems, as it considers a wide set of existing tools to evaluate information quality in the Web context, namely site analyzers, traffic analyzers, port scanners, performance monitoring systems, Web mining tools and survey tools to generate opinion-based user feedback. Several *information quality criteria* (in our terminology, dimensions and metrics, as seen in Section 3.3) can be measured with the help of these tools. IQM provides systematic sequential steps to match information quality criteria with measurement tools.

3.7. Summary Comparison of Methodologies

The detailed comparison of methodologies discussed in the previous sections clearly indicates that methodologies tend to focus on a subset of DQ issues. The broad differences in focus across methodologies can be recognized at a glance by classifying methodologies into four categories, as shown in Figure 5:

—*complete methodologies*, which provide support to both the assessment and improvement phases, and address both technical and economic issues;

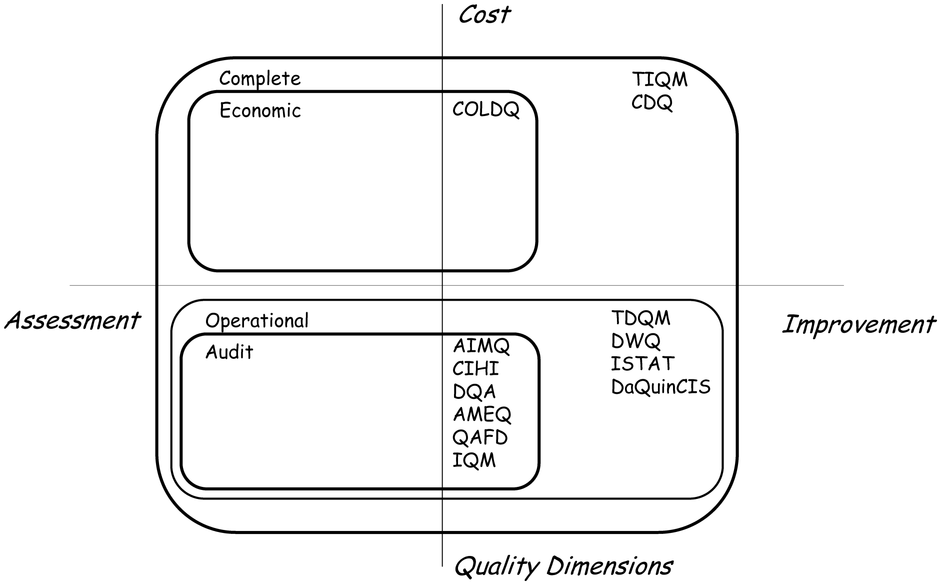


Fig. 5. A classification of methodologies.

—*audit methodologies*, which focus on the assessment phase and provide limited support to the improvement phase;

—*operational methodologies*, which focus on the technical issues of both the assessment and improvement phases, but do not address economic issues.

—*economic methodologies*, which focus on the evaluation of costs.

From a historical perspective, there exists a correlation between quality dimensions and the evolution of ICT technologies. First-generation information systems (in the '50s and '60s of the past century) were monolithic, as their technological architecture consisted of a single mainframe and a single database. Information flows were simple and repetitive and most errors were caused by incorrect data entry. The main large-scale applications were census management and medical data analysis, and data quality focused on accuracy, consistency, completeness, and time-related dimensions, consistent with our definitions in Section 2.3 and also Catarci and Scannapieco [2002]. The most critical issues with data quality management were error localization and correction in data sources, and record linkage between new data sources and pre-existing data bases.

The evolution of information systems from monolithic to network-based has caused a growth of the number of data sources in both size and scope and, consequently has significantly increased the complexity of data quality management. DQ methodologies have started to focus on new quality dimensions, such as the completeness of the data source, the currency of data, and the consistency of the new data sources compared to the enterprise database. With the advent of the Web, data sources have become difficult to assess and control over time. At the same time, searching and navigating through the Web is potentially unlimited. As a consequence of this fast evolution, methodologies have started to address new quality dimensions, such as *accessibility* and *reputation*. *Accessibility* measures the ability of users to access data, given their culture, physical status and available technologies, and is important in cooperative and network-based information systems. *Reputation* (or *trustworthiness*) is a property of data sources measuring their ability to provide correct information and is particulary relevant in Web-based and peer-to-peer systems, as further discussed in Section 4.

This evolution of ICT and the consequent growing complexity of DQ is a fundamental reason why methodologies have specialized on a subset of DQ issues, as shown in Figure 5. The majority of methodologies belong to the audit category and focus on the technical issues of the assessment phase. To some extent, this is related to the need for assessing data quality as part of improvement activities in order to evaluate the effectiveness of the improvement techniques. However, the assessment of quality also raises novel and complex DQ issues that are particularly interesting from a scientific perspective. For example, the definition of metrics for dimensions such as completeness and accuracy is heavily grounded on database theory and broadens the scope of traditional database design issues, such as view integration and schema normalization. In contrast, the economic issues of assessment and improvement require an empirical approach. The experimentation of techniques in current real-world contexts can be challenging and only a few methodologies focus on providing empirical evidence on the economics of DQ.

It should be noted that audit methodologies are more accurate than both complete and operational methodologies in the assessment phase. First of all, they are more detailed as to how to select appropriate assessment techniques, by providing more examples and related contextual knowledge. Second, they identify all types of issues, irrespective of the improvement techniques that can or should be applied. AIMQ and QAFD methodologies, for instance, describe in detail how objective and subjective assessments can be performed and provide guidelines to interpret results. DQA discusses the operating definitions that can be used to measure the different DQ dimensions, to evaluate aggregate measures of DQ for databases and, more recently, data sources.

Operational methodologies focus DQ assessment on identifying the issues for which their improvement approach works best. One of the main contributions is the identification of a set of relevant dimensions to improve and the description of a few straightforward methods to assess them. For example, TDQM is a general-purpose methodology and suggests a complete set of relevant dimensions and improvement methods that can be applied in different contexts. The completeness decreases as methodologies focus on particular contexts. For example, DWQ analyzes the data warehouse context and defines new quality dimensions tailored to the architecture of a data warehouse. The list of relevant dimensions represents an important starting point for the improvement process, since it supports companies in the identifications of the DQ issues affecting their datawareouse. Note that the assessment procedures are described more precisely in operational methodologies that focus on a specific context, rather than in general-purpose methodologies. Thus, the specialization of operational methodologies reduces their completeness and applicability if compared with complete methodologies, but increases the efficiency of the proposed techniques.

Furthermore, operational methodologies are more efficient when they focus on a particular DQ issue, especially in the application of data-oriented techniques. For example, the ISTAT methodology focuses on localization data, that relate the addresses within an administrative territory with the personal data of resident people. Such data are most frequently exchanged across agencies. In this domain, the ISTAT methodology compares the different heterogeneous formats adopted for localization data, the data owners of the different domain attributes, such as the ZIP code, and proposes a unified format and a new approach to data exchange based on the XML markup language.

As a second example, if DQ issues are related to the accuracy and completeness of personal data, improvement methodologies can be more straightforward in targeting record linkage techniques; this is the case of the DaQuinCIS and ISTAT methodologies, that use record linkage to integrate different sources, by providing domain-specific similarity algorithms for the discovery of duplicate records. For example, deduplication of names of streets is performed in bilingual regions, such as the Alto Adige region in Italy, adopting similarity functions specialized to paradigmatic errors such as imputing "u" instead of "*ü*".

Another issue addressed in operational methodologies is the governance of the DQ process and related risk management and feasibility problems. In this respect, the original contribution of the ISTAT methodology versus all other methodologies, is its focus on the large-scale application to all the central administrations of a country, including peripheral organizational units, distributed over the territory, but hierarchically dependent on central agencies. Usually, such a group of administrations has critical characteristics, such as (1) a high complexity, in terms of interrelations, processes, and services in which different agencies are involved, due to the fragmentation of competencies; (2) a high level of autonomy, which makes it difficult to enforce common rules; (3) a high heterogeneity of meanings and representations of data and data flows; and (4) large overlaps among heterogeneous records and objects. Improving DQ in such a complex structure is usually a very risky and costly activity. This is the reason why the main goal of the ISTAT methodology is to achieve feasibility. Therefore, attention is primarily focused on the most common type of data exchanged between agencies, namely, address data. In order to reduce the complexity of the problem, an assessment is first performed on the most relevant data bases and data flows to discover the most critical areas. The methodology mentions an experience where the less accurate records among address data concern regions where street names are bilingual. In a second phase, more in-depth assessment activities are performed on the local databases owned by different administrations under their responsibility, using common techniques and tools that are centrally provided. Finally, data exchanges among agencies are centrally planned and coordinated.

Complete methodologies are extremely helpful in providing a comprehensive framework to guide large DQ programs in organizations that process critical data and attribute to DQ a high strategic priority, such as banks and insurance companies. On the other hand, they show the classical tradeoff between the applicability of the methodology and the lack of personalization to specific application domains or technological contexts. Being high-level and rather context independent, complete methodologies are only marginally affected by the evolution of ICT technologies and, over time have been revised to encompass the variety of data types, sources, and flows that are part of modern information systems. For example, the IP-MAP model of TDQM has evolved to IP-UML in order to manage the growing complexity of systems in terms of processes and actors. However, its role and use within the overall framework of TDQM has not changed significantly.

Economic methodologies complement other methodologies and can be easily positioned within the overall framework provided by any complete methodology. Most audit and improvement methodologies have a cost evaluation step (see Table XI). However, they mostly focus on the cost of DQ initiatives, while a complete cost-benefit analysis should also consider the cost of doing nothing—the cost of poor data quality, which is typically of an organizational nature. Economic methodologies focus on both aspects. In particular, COLDQ focuses on the evaluation of the cost associated with poor data quality, characterizing the economic impact based on the ultimate effects of bad data. The result is called the data quality scorecard, and can be used as a tool to find the best solutions for improvement. In CDQ, the overall evaluation of the cost of poor quality is further developed to take into account the fact that the same quality improvement can be obtained with different priorities and paths and the minimization of the cost of the data quality program is the main criterion to choose among alternative improvement processes. The methodology suggests that different improvement processes can identified as paths of data- and process-driven techniques applied to the data bases, data flows and document bases involved in the improvement. Then, the costs of the different processes should be evaluated and compared, to select the minimum-cost process.

Concerning the validation of methodologies in real application contexts, we describe in the following, the most significant experiences reported in the literature. The description proceeds case by case due to the dispersed nature of the experiences reported. In fact, the empirical validation of methodologies is often missing or based on case studies and is seldom generalized with large scale scientific experimentations, since it typically takes place in industrial contexts and is part of companies' core competencies for consulting in the field. As a consequence, process testing typically remains unpublished.

Experiences of use of early TDQM versions are reported in several U.S.A. Department of Defence (DoD) documents (see US Department of Defense [1994]). Specifically, the use of DQ tools developed over SQL scripts and programming approaches to check data quality are supported. Recently, the methodology has been the basis for a law enforcement tool [Sessions 2007]. Other applications of TDQM performed at the DoD Medical Command for Military Treatment Facilities (MTF) are reported in Wang [1998] and Corey et al. [1996]. A claimed advantage of TDQM is that based on target payoffs, critical DQ issues, and the corresponding types of data, one can effectively evaluate how representative and comprehensive the DQ metrics are and whether this is the right set of metrics. This is the reason for its extensive application in different contexts, such as insurance companies, as described in Nadkarni [2006] and Bettschen [2005]. Nowadays, there are also many contributions extending TDQM, by improving IP-MAP [Scannapieco et al. 2005; Shankaranarayanan and Wang 2007] or by proposing a TDQM based Capability Maturity Model [Baskarada et al. 2006]. TIQM is a professional methodology. A significant variety of real-life examples and case studies is discussed in English [1999], ranging from customer relationship management to telemarketing and healthcare. A large number of DQ tools are also compared. The managerial principles discussed in the book and the numerous success stories reported in the related Internet site www.infoimpact.com, provide examples of relevant professional experiences, especially in the area of leveraging data quality for cost reduction, improvement of information value, and business effectiveness.

CIHI has been conceived and applied for many years in organizing, maintaining, and improving over 20 health administrative databases, many of which are person oriented and population based [Long and Seko 2005]. Evidence is mounting across Canada that reflects CIHI's impact in supporting effective health care management and developing public policy [Chapman et al. 2006]. CIHI is also used as a reference model in other countries; in fact CIHI together with TDQM have also been used to establish a data quality strategy for the Ministry of Health in New Zealand [Kerr and Norris 2004]. After long-term application, the authors provide success stories that show that the evaluation process when using CIHI appears to have been successful in meeting the primary objective of identifying and ranking the most critical issues of data quality improvement. Not only does it highlight adequate areas or conversely, target problematic areas within a database, it also appears to facilitate the seemingly overwhelming task of understanding the state of data quality for numerous data holdings. Furthermore, the evaluation scores can be used to describe data quality for a generic institution. It appears that strategic, corporate-wide planning might be facilitated when evaluation results are summarized across the holdings. It is also noted that, even though the number of evaluations is still low, numerous database improvements have already been implemented and many of these improvements might not have been detected otherwise. The main limitations experienced with CIHI are that the measurement properties of the evaluation process are not yet known and only preliminary data were available at the time of the study.

The authors of QAFD claim that their work is the result of many years of experience in the financial data quality area and that data quality analysis techniques in QADF are easy to develop and do not require the use of expensive solution packages. They also provide a real-case scenario, although, for security reasons, they do not disclose all available evidence.

In Batini and Scannapieco [2006], a large-scale experience of the application of CDQ is reported, referring to the reorganization of Government to Business (G2B) relationships in Italy. The interactions with government are needed for several business events, such as starting a new business and evolving a business, which includes variations in legal status, board composition, senior management, and number of employees. In their interactions with businesses, agencies manage information common to all businesses, typically official name, headquarters, branch addresses, main economic activity. Since every business independently reports to each agency, the copies have different levels of accuracy and currency. CDQ has been applied in this context leading to, (1) the execution of record-linkage activities with the goal of linking the diverse business identifiers in the registries of different administrations; and to (2) the reengineering of the G2B relationship. In the reengineered scenario, businesses interact with only one administration, that is in charge of sending the same piece of information to other administrations. Considering a three-year period limited to costs and savings related to data quality, benefits from the application of CDQ have been estimated around 600 million Euros. This provides evidence of the validity of the methodology, in which both data- and process-driven activities are considered, and the data quality program is chosen by optimizing the cost/quality ratio. Other applications of CDQ are reported in Basile et al. [2007], where a tool adopting CDQ as reference methodology has been used for the evaluation, measurement, and mitigation of the operational risks related to the Basel II process, leading, in some cases, to significant yearly savings.

4. CONCLUSIONS AND OPEN ISSUES

In this article we have addressed the issue of methodologies for data quality, describing and comparing thirteen of them. The whole DQ research field is currently evolving, and cannot be considered mature. Methodologies for data quality measurement and improvement are evolving in several directions: (1) considering a wider number of data types, moving from data quality to information quality, (2) relating data quality issues more closely to business process issues; and (3) considering new types of information systems, specifically Web and P2P information systems. We discuss the three areas separately and discuss open problems.

4.1. From Data Quality to Information Quality

Our previous analyses highlight the fact that information quality issues are only marginally addressed in the area of semistructured data, unstructured data, and multimedia. More precisely, methods and techniques have long been developed for such types of data in separate research areas, such as natural language understanding for documents, with scarce or no cross-fertilization with the area of data quality.

The limited number of research contributions focusing on semistructured and unstructured data in the DQ domain are a consequence of the tight historical relationship between DQ and database design. Even complete DQ methodologies are biased by an underlying focus on large collections of structured data, which still represent the most mature information resource in most organizations [Batini et al. 2008]. The interest in semistructured and unstructured data as organizational resources is more recent. Knowledge management, Web preservation, and geographical information systems represent important research fields where DQ techniques for unstructured and semistructured data are currently investigated [Batini and Scannapieco 2006]. Developing DQ techniques for semistructured and unstructured data in these fields requires a higher degree of interdisciplinarity, which, in turn, may involve an additional delay.

4.2. Data Quality and Process Quality

The relationship between data quality and process quality is a wide area of investigation, due to the relevance and diversity of characteristics of business processes in organizations. The different impacts of data quality at the three typical organizational levels, namely the operational, the tactical, and the strategic levels, are analyzed in Redman [1998], reporting interviews and outcomes of several proprietary studies. Data quality and its relationship with the quality of services, products, business operations, and consumer behavior is investigated in very general terms in Sheng and Mykytyn [2002] and Sheng [2003], where generic propositions such as "the information quality of a firm is positively related to the firm's performance" are substantiated with empirical evidence. The problem of how improving information production processes positively influences data and information quality is also analyzed in English [2002].

A few papers address more specific issues, and, consequently, present more concrete results. Vermeer [2000] examines the issue of electronic data interchange (EDI), which concerns the exchange of data among organizations using standard formats, and its influence on the efficiency and effectiveness of business processes. EDI and, more generally, markup languages, are seen as a DQ technology enabler, since they potentially reduce paper handling activities, data-entry errors, and data-entry functions.

The impact of data quality on the effectiveness of decision-making activities is investigated in Raghunathan [1999], with a focus on the accuracy dimension. The analysis shows that the effectiveness of decisions improves with a higher data quality only if the decision maker has knowledge about the relationship among problem variables, while it may degrade in the opposite case.

The influence of data quality in extreme process conditions, such as disasters and emergencies, is discussed in Fisher and Kingma [2001]. Flaws in accuracy, completeness, consistency, and timeliness are considered with reference to critical situations, such as the US Navy cruiser Vincennes firing at an Iranian Airbus, that brought 290 people to their deaths.

The role of information in the supply chain is considered in Dedeke [2005], where the *quality robustness* of an information chain is proposed to measure the ability of the information production process to also build the final information product in case of threats that cause information distortion, transformation variabilities and information failures. A methodological framework called *process quality robustness design* is proposed as a framework for diagnosing, prescribing, and building quality into information chains.

4.3. Data Quality and New Types of Information Systems

With the evolution of technology, the nature of information systems is changing and new types of information systems are emerging. We analyze issues related to Web and P2P information systems.

Concerning Web information systems, methodologies address several problems: (1) the quality of unstructured data and, in particular, documents; (2) new types of quality dimensions, such as accessibility.

Pernici and Scannapieco [2003] propose a model that associates quality information with Web data, namely with each item in a Web page, with pages, and with groups of pages. This model is applied within a methodology for data quality design and management in Web information systems. The authors discuss how to enrich methodologies for web information system design (such as Mecca et al. [1998] and Isakowitz et al. [1995]) with additional steps specifically devoted to data quality design. Several dimensions are considered, such as volatility, completability, and semantic and syntactic accuracy.

The quality of Web documents is of increasing relevance, since the number of documents that are managed in Web format is constantly growing. Several studies (e.g., Rao [2003]) have shown that 40% of the material on the net disappears within one year, while a further 40% is modified, leaving only 20% in its original form. Other studies [Lyman and Varian 2003] indicate that the average lifetime of a Web page is 44 days and the Web changes completely about four times in a year. As a consequence, the preservation of Web data becomes more and more crucial. The term *Web preservation* indicates the ability to prevent the loss of information in the Web, by storing all significant versions of Web documents.

Cappiello et al. [2003a] propose a methodology to support the preservation process over the entire life cycle of information, from creation, to acquisition, cataloguing, storage, and access. The main phases of the methodology are summarized in the following.

(1) Each time a new page is published, data are associated with metadata, describing their quality, in terms of accuracy, completeness, consistency, currency, and volatility, as defined in Section 2.3.
(2) In the acquisition phase, the user specifies acceptable values for all quality dimensions. If new data satisfy quality requirements, they are incorporated into an archive. Otherwise, data are returned to their owner and are not catalogued until their quality is satisfactory.
(3) In the publishing stage, when a new page replaces an old Web page, the volatility of old data is evaluated. If old data are still valid, data are not deleted and are associated with a new URL.

Assessment methodologies for evaluating specific qualities of Web sites are proposed in Atzeni et al. [2001], Mecca et al. [1999], and Fraternali et al. [2004]. Atzeni et al. [2001] is specifically focused on *accessibility*, as defined in Section 2.3, evaluated on the basis of a mixed quantitative/qualitative assessment. The quantitative assessment activity checks the guidelines provided by the World Wide Web Consortium in [World Wide Web Consortium www.w3.org/WAI/]. The qualitative assessment is based on experiments performed with disabled users. Fraternali et al. [2004] focus on the *usability* of the site and propose an approach based on the adoption of *conceptual logs*, which are Web usage logs enriched with metadata inferred from the conceptual schema of the Web site. While more traditional measures of the quality of a Web site are based on the hypertext representation of information, in this approach new indicators are proposed based on a conceptual representation of the site.

P2P systems are completely open and raise a need to assess and filter data. Trust-related quality dimensions in peer-to-peer systems are still an open issue. A possible solution is to rely on the reputation (or trustworthiness) of each peer. As an example, the *rating service* of DaQuinCIS (see Figure 4) associates trust values with each data source in the information system. The rating service is centralized and is supposed to be performed by a third-party organization. Trust values are used to determine the reliability of the quality evaluations made by organizations. Metrics are based [De Santis et al. 2003] on complaints the users of the data sets produced by each peer. The metrics proposed in Gackowski [2006] consider as likely accurate, the data sources that exchange data with accurate or high-quality data sources. These metrics also include a temporal component that takes into account data source past accuracy. Other proposals suggest investigating techniques that can improve the credibility of data values [Gackowski 2006].

4.4. Further Open Issues

Further open problems in DQ methodologies concern:

(1) the identification of more precise statistical, probabilistic, and functional correlations among data quality and process quality, with a focus on (1) the empirical validation of the models; and (2) the extension of the analysis to a wider set of dimensions and to specific types of business processes;

(2) the validation of methodologies; Often, a methodology is proposed without any large-scale specific experimentation and with none or only a few, supporting tools. There is a lack of research on experiments to validate different methodological approaches and on the development of tools to make them feasible;

(3) the extension of methodological guidelines to a wider set of dimensions, such as performance, availability, security, accessibility, and to dependencies among dimensions. An example of a dependency among currency and accuracy is the following: if an item is not current it is also inaccurate in 70% of the cases. Knowledge on dependencies can be acquired with data mining techniques. Further, dependencies can be analyzed with statistical techniques, providing new knowledge for improving the efficiency and effectiveness of the improvement process [De Amicis et al. 2006].

(4) In Web information systems and in data warehouses, data are managed at different aggregation levels. *Quality composition* should be investigated to obtain aggreate quality information from the quality metrics associated with elementary data.

APPENDIXES

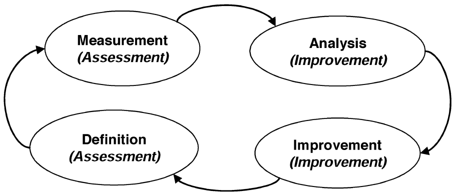
A. A SUMMARY DESCRIPTION OF METHODOLOGIES

In this appendix we provide a description card of each methodology. The description cards have a common structure composed of three parts summarizing: (1) the phases of each methodology and their mutual dependencies and critical decisions, expressed with a diagram and with a textual description, including inputs, outputs, and steps; (2) a general description highlighting the focus of each methodology and original contributions to the data quality assessment and improvement process; and (3) detailed comments discussing the applicability of each methodology.

In the diagrammatic and in the textual descriptions of methodologies, we report the terms used in the methodology, together with the standard terms we have used in Section 2.1 (reported in parentheses).

A.1. The TDQM (Total Data Quality Management) Methodology

General description. The TDQM methodology was the first general methodology published in the data quality literature [Wang 1998]. TDQM is the outcome of academic research, but has been extensively used as a guide to organizational data reengineering initiatives. The fundamental objective of TDQM is to extend to data quality, the principles of Total Quality Management (TQM) [Oakland 1989]. In operations management,



Phase 1: **Definition** .
Input → Information product
Data Analysis: Definition of the characteristics at two levels: (1) high level: description of functionalities for information consumers
(2) low level description of the basic units and components of the information product and their relationships
DQ Requirements Analysis*:* Definition of the IQ requirements from the perspective of IP suppliers, manufacturers, consumers and managers.
Process Modeling*:* Definition of the Information Manufacturing System
Output → Logical and physical design of the Information Product with the necessary quality attributes.
A quality entity-relationship model that defines the IP and its IQ requirements
An information manufacturing system that describes how the IP has been produced

Phase 2: **Measurement**
Input→ IQ dimensions
Measurement of quality: Definition of the IQ metrics
Output → IQ problems

Phase 3: **Analysis**
Input → IQ problems
Identification of the causes of errors: Determination root causes of discrepancies
Output → Actions for improving data quality

Phase 4: **Improvement**
Input → IQ metrics
Selection of strategies and techniques: Identification of key areas for improvement
Output → IQ improvement techniques

Fig. 6. Phases of TDQM.

TQM has shifted the focus of reengineering activities from efficiency to effectiveness, by offering methodological guidelines aimed at eliminating discrepancies between the output of operating processes and customers' requirements. Given requirements, reengineering must start from modeling operating processes. Consistent with these tenets, TDQM proposes a language for the description of information production (IP) processes, called IP-MAP [Shankaranarayan et al. 2000]. IP-MAP has been variously extended, towards UML and also to support organizational design. IP-MAP is the only language for information process modeling and represents a de facto standard. Practical experiences with TDQM are reported, for example, in Kovac and Weickert [2002].

Detailed comments. TDQM's goal is to support the entire end-to-end quality improvement process, from requirements analysis to implementation. As shown in Figure 6(a) TDQM Cycle consists of four phases that implement a continuous quality improvement process: definition, measurement, analysis, and improvement.

The roles responsible for the different phases of the quality improvement process are also defined in TDQM. Four roles are distinguished: *information suppliers*, which create or collect data for the IP, *information manufacturers*, which design, develop, or maintain data and related system infrastructure, *information consumers*, which use data in their work, and *information process managers*, which are responsible for managing the entire information production process throughout the information life cycle.

TDQM is comprehensive also from an implementation perspective, as it provides guidelines as to how to apply the methodology. In applying TDQM, an organization must: (a) clearly understand the IPs; (b) establish an *IP team* consisting of a senior executive as the TDQM champion, an *IP engineer* who is familiar with the TDQM methodology, and members who are information suppliers, manufacturers, consumers, and IP managers; (c) teach IQ assessment and IQ management to all the IP constituencies; and (d) institutionalize continuous IP improvement.

TDQM relies on the information quality literature for IQ Criteria and IQ improvement techniques. In particular, it explicitly refers to Wang and Strong [1996] for the data quality dimensions specification. TDQM relates quality issues to corresponding improvement techniques. However, in the recent literature no industry-specific technique is referred and no support is offered to specialize general quality improvement techniques.

A.2. The DWQ (Data Warehouse Quality) Methodology

General description. The DWQ methodology has been developed within the European Data Warehouse Quality project [Jeusfeld et al. 1998]. This methodology studies the relationship between quality objectives and design options in data warehousing. The methodology considers the subjectivity of the quality concept and provides a classification of quality goals according to the stakeholder group that pursues these goals. On the other hand, they consider the diversity of quality goals and define corresponding metadata.

Detailed comments. The DWQ methodology states that data warehouse metadata should account for three perspectives: a conceptual business perspective focusing on the enterprise model, a logical perspective focusing on the data warehouse schema, and a physical perspective representing the physical data transport layer. These perspectives correspond to the three traditional layers of data warehousing, namely sources, data warehouse, and clients. The methodology associates with each perspective, a corresponding metadata view called *Quality Measurement*.

From a data quality perspective, four main phases characterize the methodology: definition, assessment, analysis, and improvement (see Figure 7). One of the main contributions provided by this methodology is the classification of data and software quality dimensions in the data warehouse context. Three categories of data and metadata are defined:

—*Design and administration quality*: the former refers to the ability of a model to represent information adequately and efficiently, while the latter refers to the way the model evolves during the data warehouse operation.

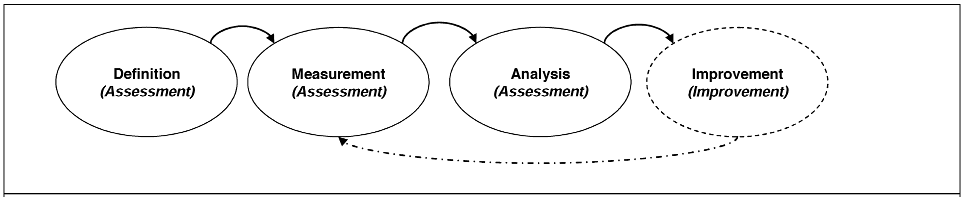
—*Software implementation quality*: the quality dimensions of the ISO 9126 standard are considered, since software implementation is not a task with specific data warehouse characteristics.

—*Data usage quality*: it refers to the dimensions that characterize the usage and querying of data contained in the data warehouse.

For each dimension contained in the listed classes, suitable measurement methods are identified. The list of these methods together with the relevance degree associated with each dimension by stakeholders are the input for the effective measurement step. In the quality assessment phase, there is the storage of the following information about each data quality dimension: (1) quality requirements—an interval of expected values; (2) the achieved quality measurement; (3) the metric used to compute a measurement; (iv) causal dependencies to other quality dimensions. Information about dependencies among quality dimensions is used to trace and analyze quality problems. The identification of critical areas is the last step analyzed in the methodology. Indeed, it only mentions the improvement phase but does not contain constructive knowledge about how to improve the quality of a data warehouse.

A.3. The TIQM (Total Information Quality Management) Methodology

General description. The TIQM methodology [English 1999] has been designed to support data warehouse projects. The methodology assumes the consolidation of



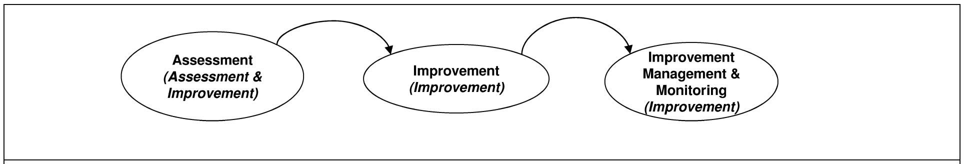
Phase 1: **Definition**
Input → Data Warehouse project and context information, operational databases, stakeholders' perspective
Data Analysis: Identification of relevant data quality dimensions in the data warehouse context and identification of the relations between data quality dimensions and data warehouse objects
DQ Requirements Analysis: Definition of the House of quality in which stakeholders assign weights to quality dimensions in order to express their relevance
Output → List of the quality dimensions and related assessment methods, quality requirements

Phase 2: **Measurement**
Input→ List of the quality dimensions and related assessment methods
Measurement of quality: Hierarchical Quality Assessment and identification of dependencies among data quality dimensions
Output → DQ values, dependencies among data quality dimensions

Phase 3: **Analysis**
Input → DQ values, dependencies among data quality dimensions, DQ Requirements
Identification of critical areas: Comparison between data quality values and data quality requirements
Output → List of dimensions that need improvement, effects of dependencies

Phase 4: **Improvement (only mentioned)**
Input → List of dimensions that need improvement, effects of dependencies
Output → Improved data

Fig. 7. Phases of DWQ.


Assessment
(Assessment & Improvement)
Improvement
(Improvement)
Improvement Management & Monitoring
(Improvement)

Phase 1: **Assessment**
Input → Information quality criteria and requirements
Data Analysis: Identify information groups and stakeholders
DQ requirements analysis: assess consumer satisfaction
Measurement of quality: Identify data validation sources, extract random samples of data, and measure and interpret data quality
Evaluation of costs*:* Identify business performance measures, calculate non quality costs, and evaluate benefits
Output → Information quality assessment

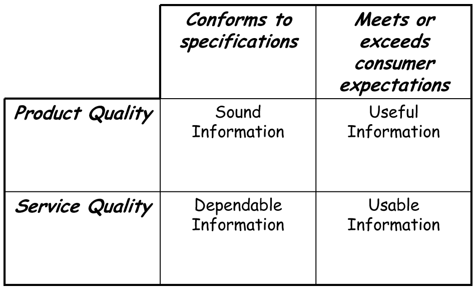
Phase 2: I**mprovement**
Input→ Information quality assessment results
Identification of the causes of errors: Analyze data defect types
Design of data improvement solutions*:* Standardize data, correct, complete, match, transform and consolidate data
Process control
Process redesign
Output → Improvement solutions

Phase 3: **Improvement Management & Monitoring**
Input → Improvement directives from general improvement
Improvement management:: Discover the organization's level of maturity, Create a vision for information quality improvement, Conduct customer satisfaction surveys of the information stakeholders, Select a small and payoff area to conduct a pilot project, Define the information value chain, Define information stewardship, Analyze the systematic barriers to DQ and recommend changes
Improvement monitoring: Check effectiveness of improvement and establish a regular mechanism of communication and education with managers
Output → Improvement solutions

Fig. 8. Phases of TIQM.

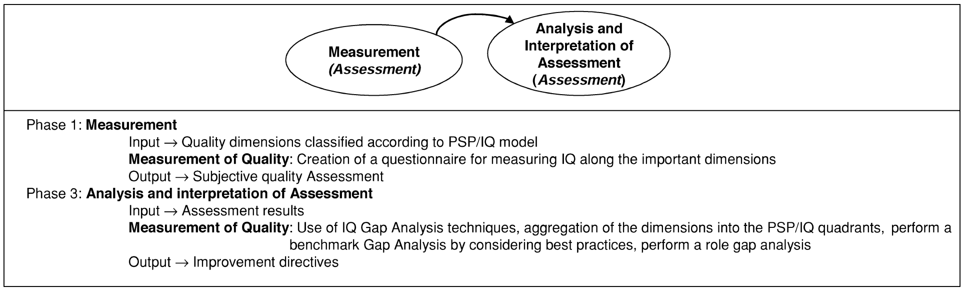
operational data sources into a unique, integrated database, used in all types of aggregations performed to build the data warehouse. This consolidation eliminates errors and heterogeneities of source databases. TIQM focuses on the management activities that are responsible for the integration of operational data sources, by discussing the strategy that has to be followed by the organization in order to make effective technical choices. Cost-benefit analyses are supported from a managerial perspective. The methodology provides a detailed classification of costs and benefits (see Section 3.4).

Detailed comments. Figure 8 shows the phases of the TIQM methodology. From the TIQM's managerial perspective, there are three main phases: assessment, improvement, and improvement management and monitoring. One of the valuable contributions of the methodology is the definition of this last phase, which provides guidelines to



	Conforms to specifications	Meets or exceeds consumer expectations
Product Quality	Sound Information	Useful Information
Service Quality	Dependable Information	Usable Information

Fig. 9. The PSP/IQ model.



Phase 1: **Measurement**

Input → Quality dimensions classified according to PSP/IQ model

Measurement of Quality: Creation of a questionnaire for measuring IQ along the important dimensions

Output → Subjective quality Assessment

Phase 3: **Analysis and interpretation of Assessment**

Input → Assessment results

Measurement of Quality: Use of IQ Gap Analysis techniques, aggregation of the dimensions into the PSP/IQ quadrants, perform a benchmark Gap Analysis by considering best practices, perform a role gap analysis

Output → Improvement directives

Fig. 10. Phases of AIMQ.

manage changes in the organization's structure according to data quality management requirements. Furthermore, the economics approach introduces cost benefit evaluation to justify data quality interventions. The goal is not only the achievement of higher data quality level, but to undertake improvement actions only if they are feasible; thus only if benefits are greater than costs.

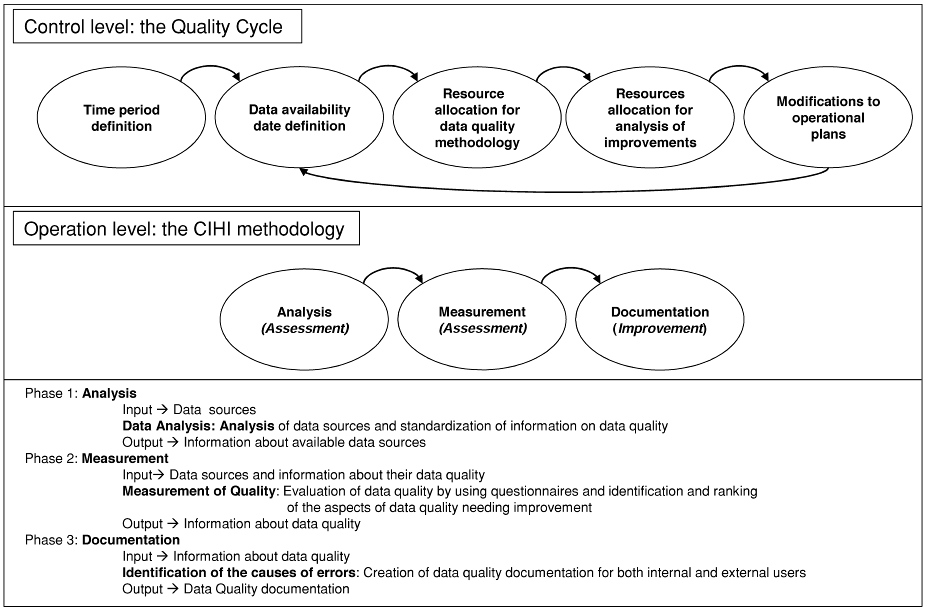
A.4. The AIMQ (A Methodology for Information Quality Assessment) Methodology

General description. The AIMQ methodology is the only information quality methodology focusing on benchmarking [Lee et al. 2002], that is an objective and domain-independent technique for quality evaluation.

The foundation of the AIMQ methodology is a 2x2 table, called the PSP/IQ model (see Figure 9), classifying quality dimensions according to their importance from the user's and manager's perspectives. The axes of the table are conformity to specifications and conformity to users' expectations. Accordingly, four classes of dimensions are distinguished (sound, dependable, useful, and usable) and quality dimensions identified in Wang and Strong [1996] are classified along these classes. Benchmarking should rank information within each class.

The PSP/IQ model is an input to the AIMQ methodology whose phases are summarized in Figure 10. The publications describing AIMQ mainly focus on the assessment activities, while guidelines, techniques, and tools for improvement activities are not provided.

Detailed comments. IQ is mainly assessed by means of questionnaires. A first pilot questionnaire is used to identify relevant quality dimensions and attributes to be benchmarked. Then, a second questionnaire addresses the dimensions and attributes previously identified in order to obtain IQ measures. Finally, these measures are compared


Control level: the Quality Cycle
Time period definition
Data availability date definition
Resource allocation for data quality methodology
Resources allocation for analysis of improvements
Modifications to operational plans
Operation level: the CIHI methodology
Analysis *(Assessment)*
Measurement *(Assessment)*
Documentation (*Improvement*)

Phase 1: **Analysis**
Input → Data sources
Data Analysis: Analysis of data sources and standardization of information on data quality
Output → Information about available data sources
Phase 2: **Measurement**
Input→ Data sources and information about their data quality
Measurement of Quality: Evaluation of data quality by using questionnaires and identification and ranking of the aspects of data quality needing improvement
Output → Information about data quality
Phase 3: **Documentation**
Input → Information about data quality
Identification of the causes of errors: Creation of data quality documentation for both internal and external users
Output → Data Quality documentation

Fig. 11. Phases of CIHI.

with benchmarks. Lee et al. [2002] provide a list of standard quality dimensions and attributes helping the definition of questionnaires.

The literature on AIMQ does not provide any description of the benchmarking database that is required for the application of the methodology. Gap Analysis techniques are advocated as a standard approach to conduct benchmarking and interpret results. In particular, two Gap Analysis techniques are suggested: Information Quality Benchmark Gaps and Information Quality Role Gaps. The former compares the quality values of an organization with those of best-practice organizations. The latter compares the information quality assessments provided by different organizational roles, that is the IS professional and the information user. IQ Role Gaps searches for discrepancies between the evaluations provided by different roles as an indication of potential quality issues. Discrepancies are associated with a direction. The direction of the gap is positive if the assessment of IS professionals is higher than the assessment of users. A large positive gap is considered dangerous, since it indicates that IS professionals are not aware of quality issues that information users have detected. If the size of the gap is small, the location of the gap should be analyzed. If the location is high, indicating high IQ, incremental improvements are most appropriate, whereas if the location is low, major improvement efforts are likely to be required.

A.5. The CIHI (Canadian Institute for Health Information) Methodology

General description. The CIHI methodology has implemented a method to evaluate and improve the quality of Canadian Institute for Health Information data [Long and Seko 2005]. In the CIHI scenario, the main issue is the size of databases and their heterogeneity. The CIHI methodology supports the selection of a subset of data to focus the quality assessment phase. It also proposes a large set of quality criteria to evaluate heterogeneity.

Detailed Comments. The CIHI Data Quality strategy proposes a two-phase approach (see Figure 11). The first phase is the definition of a Data Quality Framework, and the second is an in-depth analysis of the most frequently accessed data. The Data Quality Framework is defined in three steps: (1) standardization of data quality information; (2) development of a common strategy for data quality assessment; (3) definition of a work process for CIHI's data management that identifies data quality priorities and implements continuous data improvement procedures.

The implementation of the CIHI framework is cyclical, according to the continuous improvement approach. The following is required for a successful implementation:

—definition of the time period for the cycle;

—definition of time objectives for different quality targets;

—allocation of ad hoc resources for data quality analysis, evaluation, and documentation;

—allocation of ad hoc resources for data quality improvement.

The analysis of the most frequently used data is performed in three steps: data quality analysis, evaluation, and documentation. Documents report the quality problems detected by the data quality analysis and evaluation.

Data quality evaluation is based on a four-level hierarchical model. At the first level, 86 basic quality *criteria* are defined. These criteria are aggregated by means of composition algorithms into 24 quality *characteristics* at the second hierarchical level, and further aggregated into five quality *dimensions* at the third level; namely, accuracy, timeliness, comparability, usability, and relevance. Finally, the five dimensions are aggregated into one overall *database evaluation* at the fourth level.

The basic evaluation of the 86 data quality criteria is performed by means of questionnaires reporting criteria as items to be scored on a four-point ordinal scale as "not applicable," "unknown," "not met," or "met." Then, at each aggregation level, evaluations are validated. The validation process ensures that the interpretation and scoring of each criterion is as standard as possible.

A.6. The DQA (Data Quality Assessment) Methodology

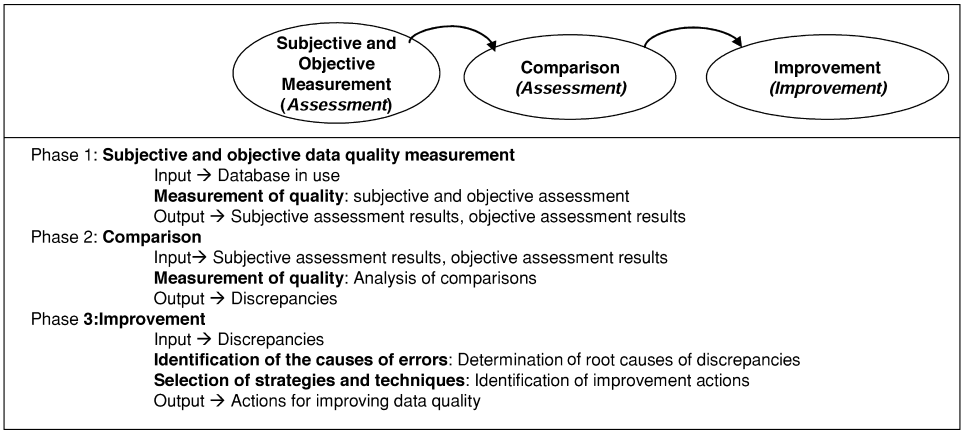
General description. The DQA methodology [Pipino et al. 2002] has been designed to provide the general principles guiding the definition of data quality metrics. In the literature, data quality metrics are mostly defined ad hoc to solve specific problems and thus, are dependent on the considered scenario. The DQA methodology is aimed at identifying the general quality measurement principles common to previous research.

Detailed comments. The classification of metrics of the DQA methodology is summarized in Figure 12. The methodology makes a distinction between subjective and objective quality metrics. Subjective metrics measure the perceptions, needs, and experiences of the stakeholders. Objective metrics are then classified into task-independent and task-dependent. The first assess the quality of data without contextual knowledge of the application, while the second are defined for specific application contexts and include business rules, company and government regulations, and constraints provided by the database administration. Both metrics are divided into three classes: simple ratio, min or max value, and weighed average.

A.7. The IQM (Information Quality Measurement) Methodology

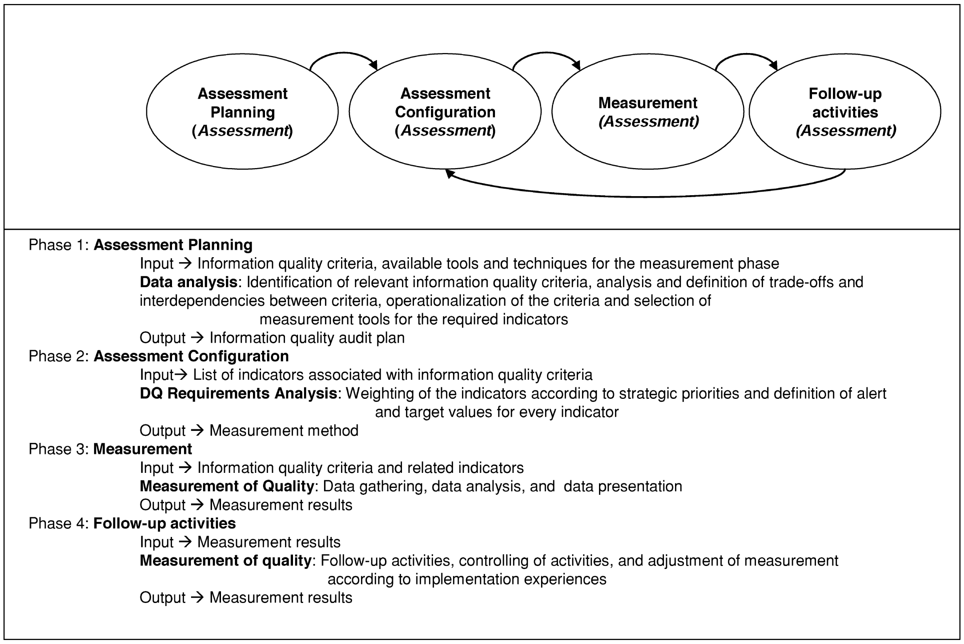
General description. The fundamental objective of the IQM methodology [Eppler and Münzenmaier 2002] is to provide an information quality framework tailored to Web data. In particular, IQM helps the quality-based selection and personalization of the tools that support Webmasters in creating, managing, and maintaining Web sites.

Detailed comments. The IQM methodology provides guidelines to ensure that software tools evaluate all the fundamental information quality dimensions. The


Subjective and Objective Measurement (*Assessment*)
Comparison *(Assessment)*
Improvement *(Improvement)*

Phase 1: **Subjective and objective data quality measurement**
Input → Database in use
Measurement of quality: subjective and objective assessment
Output → Subjective assessment results, objective assessment results
Phase 2: **Comparison**
Input→ Subjective assessment results, objective assessment results
Measurement of quality: Analysis of comparisons
Output → Discrepancies
Phase **3:Improvement**
Input → Discrepancies
Identification of the causes of errors: Determination of root causes of discrepancies
Selection of strategies and techniques: Identification of improvement actions
Output → Actions for improving data quality

Fig. 12. Phases of DQA.

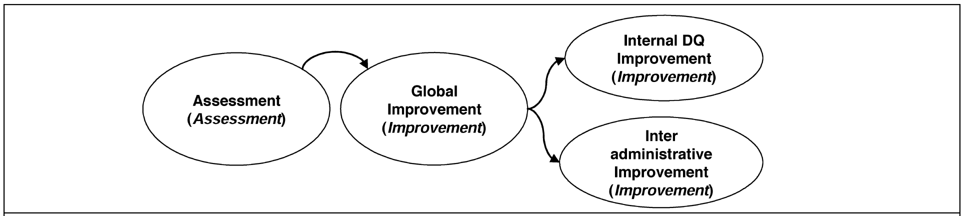

Assessment Planning (*Assessment*)
Assessment Configuration (*Assessment*)
Measurement *(Assessment)*
Follow-up activities *(Assessment)*

Phase 1: **Assessment Planning**
Input → Information quality criteria, available tools and techniques for the measurement phase
Data analysis: Identification of relevant information quality criteria, analysis and definition of trade-offs and interdependencies between criteria, operationalization of the criteria and selection of measurement tools for the required indicators
Output → Information quality audit plan
Phase 2: **Assessment Configuration**
Input→ List of indicators associated with information quality criteria
DQ Requirements Analysis: Weighting of the indicators according to strategic priorities and definition of alert and target values for every indicator
Output → Measurement method
Phase 3: **Measurement**
Input → Information quality criteria and related indicators
Measurement of Quality: Data gathering, data analysis, and data presentation
Output → Measurement results
Phase 4: **Follow-up activities**
Input → Measurement results
Measurement of quality: Follow-up activities, controlling of activities, and adjustment of measurement according to implementation experiences
Output → Measurement results

Fig. 13. Phases of IQM.

methodology provides two sets of guidelines: the *information quality framework* defining quality criteria, and the *action plan* explaining how to perform quality measurements.

The main phases of IQM methodology are reported in Figure 13. The first phase defines the measurement plan. The information quality framework is defined as a list of relevant information quality criteria identified by interviewing the information stakeholders. The framework is the input for an information quality audit that associates the information quality criteria with the methods and tools that will be used in the measurement process. Some criteria require multiple measurement methods. The IQM methodology coordinates the application of multiple measurement methods.



Phase 1: **Assessment**
Input → Information quality criteria and requirements
DQ Requirements analysis – Isolate from a general process analysis relevant qualities for address data
Identification of critical areas: Find critical areas by using statistical techniques and by analyzing sample database
Measurement of quality: Measure data quality dimensions and communicate results of assessment to single agencies
Output → Information quality assessment results
Phase 2: **Global improvement**
Input→ Information quality assessment results
Identification of the causes of errors
Design of data improvement solutions: Perform record linkage among relevant national databases
Assignment of data responsibilities: Establish a national data owner for specific fields
Process control and Process Redesign: Design improvement solutions on processes, the results of the global assessment to decide specific interventions on processes
Selection of strategies and techniques: Make or buy, and adapt, tools for most relevant techniques
Improvement management: DQ activities to deliver to agencies
Output → Improvement method
Phase 3: **Internal DQ improvement (for each agency, autonomous initiative)**
Input → Improvement directives from global improvement
Process Redesign: Standardize acquisition format and standardize internal exchange format using XML
Improvement Monitoring: Perform specific local assessments, design improvement solutions on data and processes in critical areas
Output → Improvement solutions
Phase 4: **Inter administrative improvement**
Input → Improvement directives from general improvement
Process Redesign: Standardize interadministrative flows format using XML, redesign exchange flows, using a public and subscribe event driven architecture
Output → Improvement solutions

Fig. 14. Phases of ISTAT.

A.8. The ISTAT (Italian National Bureau of Census) Methodology

General description. The ISTAT methodology [Istat 2004; Falorsi et al. 2003] has been designed within the Italian National Bureau of Census to collect and maintain high quality statistical data on Italian citizens and businesses. The fundamental issue faced by the methodology is how to guarantee the quality of data integrated from multiple databases of local Public Administrations. This issue is particularly challenging within the Italian context, where the Public Administration is organized in three geographical levels, Central, Regional and Peripheral, each managing its own data autonomously. The ISTAT methodology focuses on the most common types of data exchanged among different levels of the Public Administration, namely private data. The methodology is strongly focused on formal norms, since it is aimed at regulating data management activities in such a way that their integration can satisfy basic quality requirements.

Detailed comments. The fundamental phases of the ISTAT methodology are (see Figure 14):

—the assessment phase, that is initially performed on the central databases owned and managed by ISTAT, to detect quality issues from a data integration perspective;

—the global improvement phase, which is in charge of performing record linkage among national databases and designing the improvement solution on processes including the decision to make, buy, or adapt existing solutions;

—Improvement activities on databases owned and managed by local Administrations. These activities should be performed by local Administrations themselves, with the aid of tools and courses provided by ISTAT.

—Improvement activities that require the cooperation of multiple Administrations. These activities are typically process oriented, since they address flows of data exchanged during the execution of specific operating activities. Central databases may be required for coordination purposes. These activities are centrally planned and coordinated.

The ISTAT methodology provides a variety of simple but effective statistical techniques for quality measurement. It also provides tools for the most relevant data cleaning activities. Local Administrations are helped in tailoring tools to specific geographical or process issues. In the ISTAT methodology, data owners are defined at a high level of detail, corresponding to individual attributes, such as `MunicipalityCode`. The methodology supports the standardization of data formats and their expression in a common XML schema allowing the integration of the databases of local Administrations. Data exchanged among different Administrations are redesigned using an event-driven software architecture, based on publish and subscribe mechanisms.

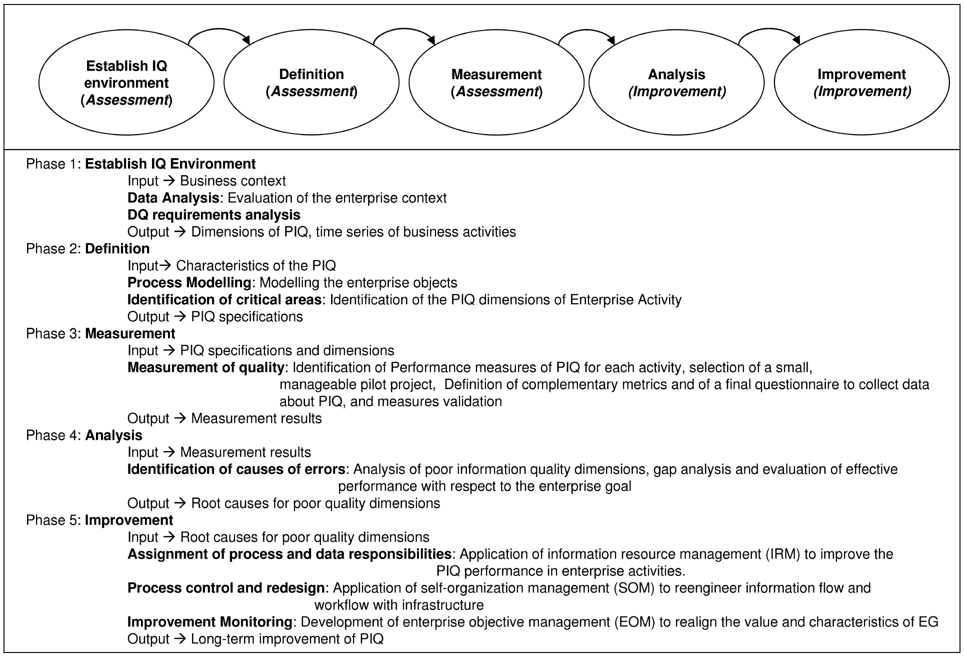
A.9. The AMEQ (Activity-based Measuring and Evaluating of Product information Quality) Methodology

General description. The main goal of the AMEQ methodology [Su and Jin 2004] is to provide a rigorous basis for Product Information Quality (PIQ) assessment and improvement in compliance with organizational goals. The methodology is specific for the evaluation of data quality in manufacturing companies, where product information represents the main component of operational databases. In manufacturing companies, the association between product information and production processes is straightforward and relatively standard across companies. The schema of product databases is also similar across different organizations. The methodology provides an approach and methodological guidelines to model both information and related production processes.

Detailed comments. The AMEQ methodology consists of five phases for measuring and improving PIQ (see Figure 15). The first phase assesses the cultural readiness of an organization, using the Information Quality Management Maturity Grid, a template to conduct interviews for key managerial roles. In this phase, the dimensions of PIQ are also defined and classified according to their relevance for different business activities. The second phase specifies the information product. Each information product is associated with a corresponding business process, modelled by means of an object-oriented approach (OOA). In the AMEQ methodology, eight types of objects are modelled: human resources, information resources, enterprise activities, resource inputs, resource processes, resource outputs, performance measures, and enterprise goals. In this phase, a model of measurement methods is also produced. The third phase focuses on the measurement activity. In the fourth phase, the IQ team should investigate the root causes for potential PIQ problems by analyzing the quality dimensions that have received a low score. Finally, the PIQ improvement phase can start. Note that for the fourth and fifth phases, AMEQ does not provide operating methods and tools, but only general guidelines.

A.10. The COLDQ (Cost-Effect Of Low Data Quality) Methodology

General description. The fundamental objective of the COLDQ methodology [Loshin 2004] is to provide a data quality scorecard supporting the evaluation of the cost-effect of low data quality. Similarly to TIQM (Section A.3), the methodology provides a detailed classification of costs and benefits (see also Section 3.4). Direct benefits are obtainable from the avoidance of poor quality costs due to the adoption of improvement techniques.


Establish IQ environment (*Assessment*)
Definition (*Assessment*)
Measurement (*Assessment*)
Analysis (*Improvement*)
Improvement (*Improvement*)

Phase 1: **Establish IQ Environment**
Input → Business context
Data Analysis: Evaluation of the enterprise context
DQ requirements analysis
Output → Dimensions of PIQ, time series of business activities
Phase 2: **Definition**
Input→ Characteristics of the PIQ
Process Modelling: Modelling the enterprise objects
Identification of critical areas: Identification of the PIQ dimensions of Enterprise Activity
Output → PIQ specifications
Phase 3: **Measurement**
Input → PIQ specifications and dimensions
Measurement of quality: Identification of Performance measures of PIQ for each activity, selection of a small, manageable pilot project, Definition of complementary metrics and of a final questionnaire to collect data about PIQ, and measures validation
Output → Measurement results
Phase 4: **Analysis**
Input → Measurement results
Identification of causes of errors: Analysis of poor information quality dimensions, gap analysis and evaluation of effective performance with respect to the enterprise goal
Output → Root causes for poor quality dimensions
Phase 5: **Improvement**
Input → Root causes for poor quality dimensions
Assignment of process and data responsibilities: Application of information resource management (IRM) to improve the PIQ performance in enterprise activities.
Process control and redesign: Application of self-organization management (SOM) to reengineer information flow and workflow with infrastructure
Improvement Monitoring: Development of enterprise objective management (EOM) to realign the value and characteristics of EG
Output → Long-term improvement of PIQ

Fig. 15. Phases of AMEQ.

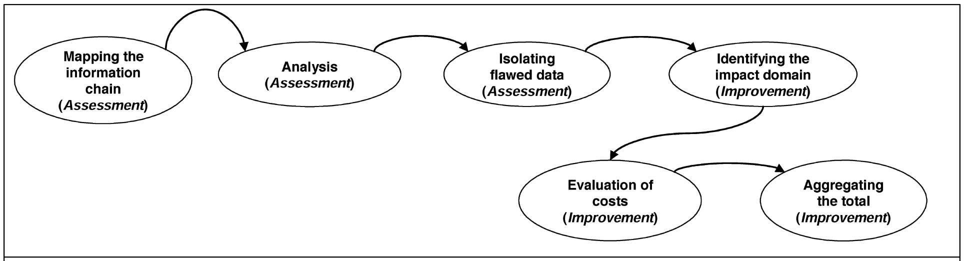
The goal is to obtain a quantitative assessment of the extent to which business processes are affected by bad information.

Detailed comments. Six interrelated phases are proposed to evaluate the cost-effect of low data quality (see Figure 16). In the first phase of the methodology, the business context is modelled by identifying two data flow models: the strategic data flow, used for decision-making, and the operational data flow, used for data processing. Both models represent a set of processing stages that describe the information flow from data supply to data consumption. Based on these models, the objective and subjective analyses of the business context are conducted. Internal and external users, employees and customers, are interviewed in order to identify flawed data. Then, errors are attributed to faulty activities in the strategic and operational models of the business context. This association between errors and activities provides the basis for cost evaluations. The COLDQ methodology provides a thorough and valuable classification of operational, tactical, and strategic economic impacts that have to be considered. Each class of costs is assigned an economic value based on contextual knowledge. Costs represent the input to the final improvement phase.

Finally, the COLDQ methodology supports cost-benefit analyses by evaluating and aggregating the cost of quality improvement projects. Methodological guidelines support the calculation of the return on investment (ROI) and break-even points of improvement initiatives.

A.11. The DaQuinCIS (Data QUality IN Cooperative Information Systems) Methodology

General description. The DaQuinCIS methodology [Scannapieco et al. 2004] addresses data quality issues in Cooperative Information Systems. Cooperation raises two context-specific data quality issues. First of all, data quality is predicated upon interorganizational trust. Second, poor data quality can hinder cooperation and, thus, has far-reaching consequences. To address the first issue, the DaQuinCIS methodology


Mapping the information chain (*Assessment*)
Analysis (*Assessment*)
Isolating flawed data (*Assessment*)
Identifying the impact domain (*Improvement*)
Evaluation of costs (*Improvement*)
Aggregating the total (*Improvement*)

Phase 1: **Mapping the information chain**.
Input → Enterprise sources
Process Modelling: Identification of steps through which the input data is converted into usable information
Output → Information chain: a map composed of the processing stages connected by directed information channels. Data flows: definition of what data are being used within the system and at which information channel source and target points the data pass through.
Phase 2: **Analysis**
Input→ Information chain maps and data flows
DQ requirement Analysis: Employees and customers interviews
Output → IQ problems
Phase 3: **Isolating flawed data**
Input → Interviews
Process Modelling: Annotation of the information chain
Output → Identification of activities attributable to data flaws
Phase 4: **Identifying the impact domain**
Input → Identification of activities attributable to data flaws
Selection of strategies and techniques: Identification of operational, tactical, and strategic impact
Output → Classification of activities related to poor data quality
Phase 5: **Evaluation of costs**
Input → Classification of activities related to poor data quality
Evaluation of costs: Association of a cost to each operational, tactical, and strategic impact
Output → Data quality scorecard matrix
Phase 6: **Aggregating the total**
Input → Data quality scorecard matrix
Evaluation of costs: Aggregation of costs
Output → Poor quality cost

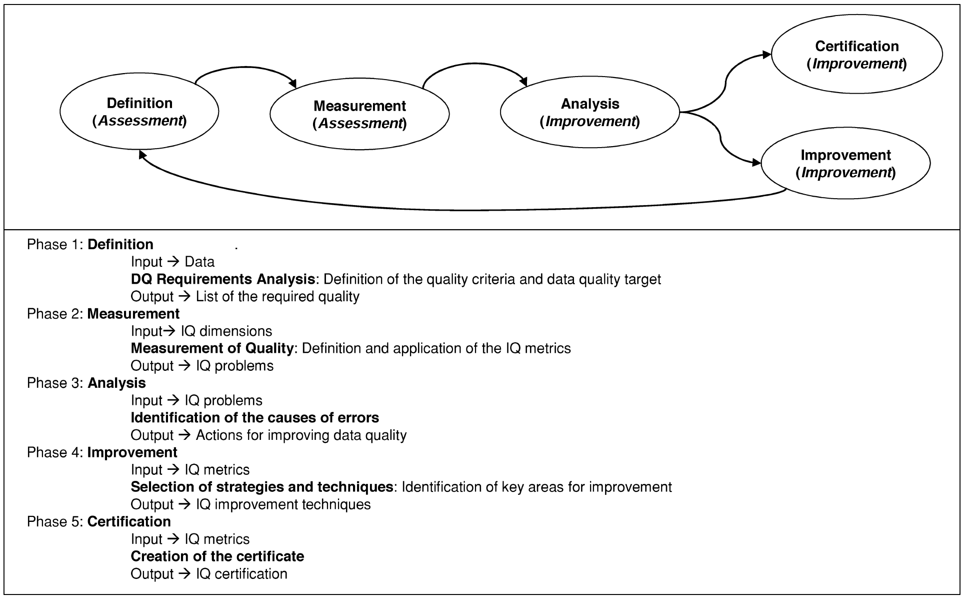
Fig. 16. Phases of COLDQ.

introduces the concept of data quality certification, which associates data with corresponding quality measures that are exchanged among organizations along with data. The second issue is addressed by providing quality-based data selection mechanisms. These selection mechanisms identify the highest-quality data among overlapping databases owned by different cooperating organizations. In this way, cooperation is leveraged to improve quality.

Detailed comments. The DaQuinCIS methodology provides an innovative model to represent data quality called *data and data quality (D^2Q)* that includes: (1) constructs to represent data, (2) a set of data quality properties, (3) constructs to represent data quality properties and (4) the associations between data and quality metadata. Source trustworthiness is included among quality properties. The value associated with this dimension is assigned by a third-party organization on the basis of several parameters, including the number of complaints made by other organizations and the number of requests issued to each source.

Figure 17 reports the fundamental phases of the DaQuinCIS methodology: *quality analysis*, *quality assessment*, *quality certification*, and *quality improvement*. The methodology is supported by an architecture composed of an internal infrastructure and an external infrastructure (see Figure 18). Methodological phases are implemented by corresponding modules of the Quality Factory that are implemented in each organization belonging to CIS. The communication among the organizations involved in the CIS is enabled by a *quality notification service* that is a publish/subscribe engine used as general message bus between the different architectural components.

In the Quality Factory, the requests from external users define data to be retrieved and corresponding quality requirements. An apposite module evaluates the quality of



Phase 1: **Definition** .
Input → Data
DQ Requirements Analysis: Definition of the quality criteria and data quality target
Output → List of the required quality
Phase 2: **Measurement**
Input→ IQ dimensions
Measurement of Quality: Definition and application of the IQ metrics
Output → IQ problems
Phase 3: **Analysis**
Input → IQ problems
Identification of the causes of errors
Output → Actions for improving data quality
Phase 4: **Improvement**
Input → IQ metrics
Selection of strategies and techniques: Identification of key areas for improvement
Output → IQ improvement techniques
Phase 5: **Certification**
Input → IQ metrics
Creation of the certificate
Output → IQ certification

Fig. 17. Phases of DaQuinCIS.

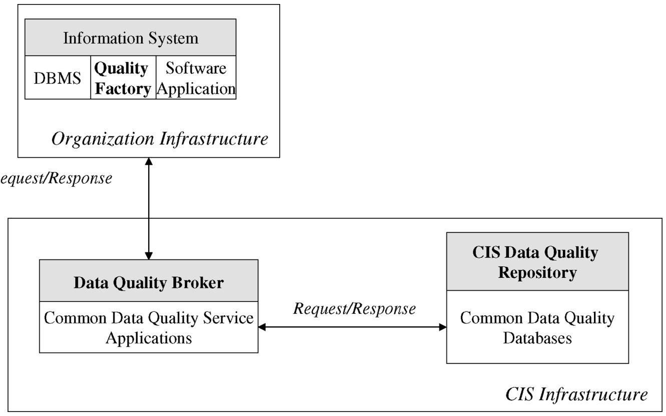
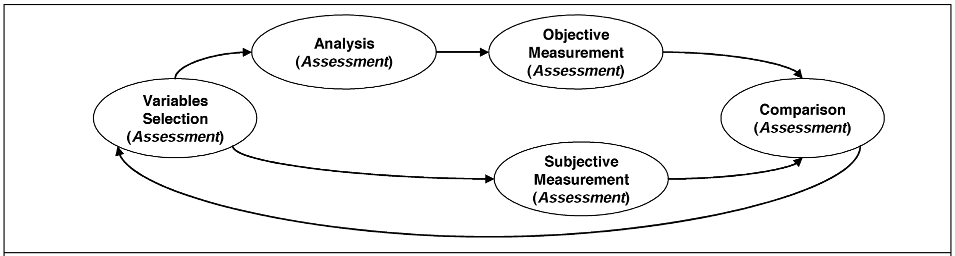


Fig. 18. The DaQuinCIS architecture.

data and compares the quality values with the quality requirements expressed by the users. If data do not satisfy the quality requirements, an alert is sent to the user. On the contrary, if values of quality are satisfactory, a quality certificate is associated with the data and sent to the user. Quality improvement is carried out by the *data quality broker*. This module, by collaborating with the *Data Quality Repository*, translates queries according to a global schema and selects data that maximize quality. A query submitted by a specific organization is issued to all organizations, specifying a set of quality requirements on requested data. Different copies of the same data received as a response to the request are reconciled and best-quality values are selected and returned to the requesting organizations.



Phase 1: **Variables selection**
Input → financial variables and the financial context
Data analysis: identification and description of the most relevant financial registry variables, classification of the variables along three classes: qualitative/categorical, quantitative/numerical, and dates
Output → classification and groups of primary variables
Phase 2: **Analysis**
Input→ classification and groups of primary variables, data quality analysis techniques, business rules, process analysis, observation
DQ Requirements Analysis: Analysis of data quality dimensions: syntactic and semantic accuracy, internal and external consistency, completeness, currency, timeliness, and uniqueness
Identification of critical areas: Application of data quality rules
Output → Identification of errors
Phase 3: **Objective measurement**
Input → Identification of errors
Measurement of quality: Definition of indices for the evaluation and quantification of the global data quality level
Output → Quantitative objective assessment
Phase 4: **Qualitative subjective measurement**
Input → Business and data quality expertise, measurement, and classification and groups of primary variables
Measurement of quality: Business expert assessment, Financial operator assessment, Data quality expert assessment, Comparison of subjective assessments
Output→ Qualitative subjective assessment
Phase 5: **Comparison**
Input→ Quantitative objective dimensions, qualitative subjective dimensions
Measurement of quality: Calculation of discrepancies
Output → Assessment and improvement

Fig. 19. Phases of QADF.

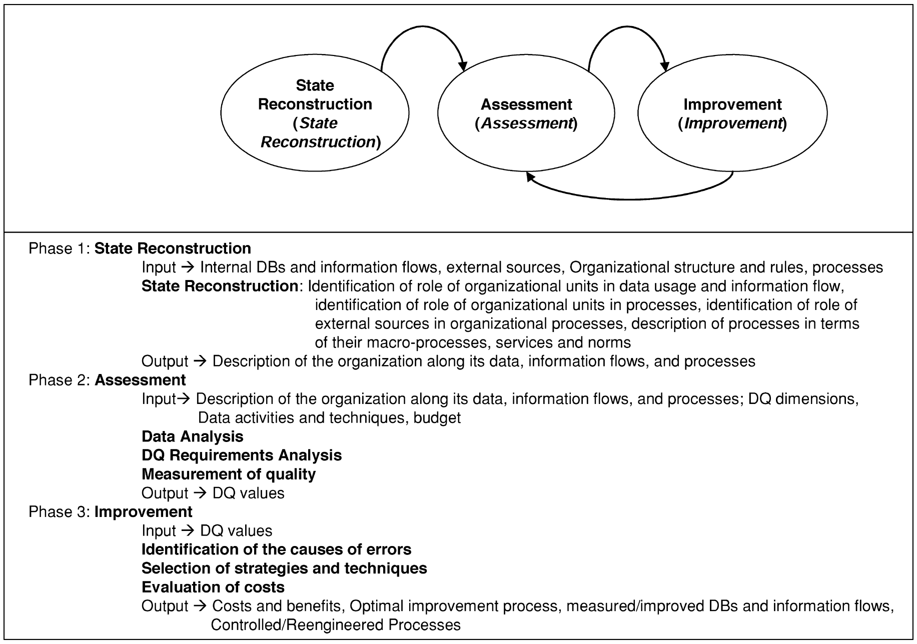
As regards the techniques used in the improvement phase, DaQuinCIS methodology proposes a new algorithm for record matching. The Record Matcher is a component of the Data Quality Broker. The Record Matcher implements a method for record matching based on the quality data exported by cooperating organizations.

A.12. The QAFD (Quality Assessment of Financial Data) Methodology

General description. The QAFD methodology [De Amicis and Batini 2004] has been designed to define standard quality measures for financial operational data and thus minimize the costs of quality measurement tools. The methodology combines quantitative objective, and qualitative subjective assessments to identify quality issues and select the appropriate quality improvement actions. Context-dependent indices, data quality rules, measurements, and strategies for quantitative and qualitative assessments are defined. Overall, it represents the only methodology for the quality assessment of financial data.

Detailed comments. The main phases of this methodology are reported in Figure 19. First, the methodology selects the most relevant financial variables. Selection is usually based on knowledge from previous assessments, according to their practical effectiveness. Variables are grouped in categories of “related issues” that similarly affect the behavior of investors and consumers and are characterized by the same risk, business, and descriptive factors.

The second phase aims at discovering the main causes of errors. The most relevant data quality dimensions are identified in this phase and data quality rules are produced. Data quality rules represent the dynamic semantic properties of variables that cannot be measured along quality dimensions.


State Reconstruction (*State Reconstruction*)
Assessment (*Assessment*)
Improvement (*Improvement*)

Phase 1: **State Reconstruction**
Input → Internal DBs and information flows, external sources, Organizational structure and rules, processes
State Reconstruction: Identification of role of organizational units in data usage and information flow, identification of role of organizational units in processes, identification of role of external sources in organizational processes, description of processes in terms of their macro-processes, services and norms
Output → Description of the organization along its data, information flows, and processes
Phase 2: **Assessment**
Input→ Description of the organization along its data, information flows, and processes; DQ dimensions, Data activities and techniques, budget
Data Analysis
DQ Requirements Analysis
Measurement of quality
Output → DQ values
Phase 3: **Improvement**
Input → DQ values
Identification of the causes of errors
Selection of strategies and techniques
Evaluation of costs
Output → Costs and benefits, Optimal improvement process, measured/improved DBs and information flows, Controlled/Reengineered Processes

Fig. 20. Phases of CDQ.

In the third phase, the objective assessment is performed based on quantitative indexes. Authors propose a mathematical model for the objective assessment resulting, in an overall ranking of data along each quality dimension.

The subjective assessment is performed in the fourth phase from three different perspectives; business experts, customers, and data quality experts. Each interviewee has to assess the quality level along each quality dimension. An overall assessment is obtained as the mean value of the subjective assessment of each class of experts.

Finally, objective and subjective assessments are compared in the fifth phase. For each dimension, the difference between the objective and subjective assessments is calculated. If the difference is positive, the objective assessment must be reconsidered to point out quality issues relevant from the experts' point of view.

A.13. The CDQ (Complete Data Quality) Methodology

General description. The CDQ methodology [Batini and Scannapieco 2006; Batini et al. 2008] is conceived to be at the same time complete, flexible, and simple to apply. Completeness is achieved by considering existing techniques and tools and integrating them in a framework that can work in both intra- and inter-organizational contexts, and can be applied to all types of data, structured, semistructured and unstructured. The methodology is flexible since it supports the user in the selection of the most suitable techniques and tools within each phase and in any context. Finally, CDQ is simple since it is organized in phases and each phase is characterized by a specific goal and set of techniques to apply.

The CDQ methodology is innovative since it provides support to select the optimal quality improvement process that maximizes benefits within given budget limits. Second, it emphasizes the initial requirements elicitation phase. In fact, the other methodologies implicitly assume that contextual knowledge has been previously gathered and modelled. The focus is on how to reach *total data quality* without providing indications as to how to use contextual knowledge. A goal of CDQ is instead to obtain a quantitative assessment of the extent to which business processes are affected by bad information.

Detailed comments. Three main phases characterize the methodology: state reconstruction, assessment, and choice of the optimal improvement process (see Figure 20). In the first phase of the methodology, the relationships among organizational units, processes, services, and data are reconstructed. These relationships are modelled by using matrixes that describe which organizational units use data and their roles in the different business processes. Furthermore, in this phase, processes are described along with their contribution in the production of goods/services and the legal and organizational rules that discipline workflows. The second phase sets new target quality levels that are needed to improve process qualities, and evaluates corresponding costs and benefits. This phase locates the critical variables affected by poor quality. Since improvement activities are complex and costly, it is advisable to focus on the parts of the databases and data flows that raise major problems. Finally, the third phase consists of five steps and is aimed at the identification of the optimal improvement process: the sequence of activities that has the highest cost/effectiveness ratio. New target quality levels are set by considering costs and benefits. Different improvement activities can be performed to reach new quality targets. The methodology recommends the identification of all the data-driven and process-driven improvement techniques for the different databases affected by poor quality. A set of mutually consistent improvement techniques constitutes an improvement process. Finally, the most suitable improvement process is selected by performing a cost-benefit analysis.

REFERENCES

ABITEBOUL, S., BUNEMAN, P., AND SUCIU, D. 2000. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers.

AIKEN, P. 1996. *Data Reverse Engineering*. McGraw Hill.

ARENAS, M., BERTOSSI, L., AND CHOMICKI, J. 1999. Consistent query answers in inconsistent databases. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. ACM, New York, 68–79.

ATZENI, P. AND ANTONELLIS, V. D. 1993. *Relational Database Theory*. Benjamin/Cummings.

ATZENI, P., MERIALDO, P., AND SINDONI, G. 2001. Web site evaluation: Methodology and case study. In *Proceedings of International Workshop on data Semantics in Web Information Systems (DASWIS)*.

BALLOU, D. AND PAZER, H. 1985. Modeling data and process quality in multi-input, multi-output information systems. *Manag. Sci. 31*, 2.

BALLOU, D., WANG, R., PAZER, H., AND TAYI, G. 1998. Modeling information manufacturing systems to determine information product quality. *Manage. Sci. 44*, 4.

BASILE, A., BATINI, C., GREGA, S., MASTRELLA, M., AND MAURINO, A. 2007. Orme: A new methodology for information quality and basel II operational risk. In *Proceeedings of the 12th International Conference of Information Quality, Industrial Track*.

BASILI, V., CALDIERA, C., ROMBACH, H. 1994. Goal question metric paradigm.

BASKARADA, S., KORONIOS, A., AND GAO, J. 2006. Towards a capability maturity model for information quality management: a tdqm approach. In *Proceedings of the 11th International Conference on Information Quality*.

BATINI, C., CABITZA, F., CAPPIELLO, C., AND FRANCALANCI, C. 2008. A comprehensive data quality methodology for Web and structured data. *Int. J. Innov. Comput. Appl. 1*, 3, 205–218.

BATINI, C. AND SCANNAPIECO, M. 2006. *Data Quality: Concepts, Methodologies and Techniques*. Springer Verlag.

BERTOLAZZI, P., SANTIS, L. D., AND SCANNAPIECO, M. 2003. Automatic record matching in cooperative information systems. In *Proceedings of the ICDT International Workshop on Data Quality in Cooperative Information Systems (DQCIS)*.

BETTSCHEN, P. 2005. Master data management (MDM) enables IQ at Tetra Pak. In *Proceedings of the 10th International Conference on Information Quality*.

Bilke, A., Bleiholder, J., Böhm, C., Draba, K., Naumann, F., and Weis, M. September 2005. Automatic data fusion with HumMer. In *Proceedings of the VLDB Demonstration Program*.

Bovee, M., Srivastava, R., and Mak, B. September 2001. A conceptual framework and belief-function approach to assessing overall information quality. In *Proceedings of the 6th International Conference on Information Quality*.

Buneman, P. 1997. Semi-structured data. In *Proceedings of the 16th ACM Symposium on Principles of Database Systems (PODS)*.

Calì, A., Calvanese, D., De Giacomo, G., and Lenzerini, M. 2004. Data integration under integrity constraints. *Inform. Syst. 29*, 2, 147–163.

Calvanese, D., De Giacomo, D., and Lenzerini, M. 1999. Modeling and querying semi-structured data. *Network. Inform. Syst. J. 2*, 2, 253–273.

Cappiello, C., Francalanci, C., and Pernici, B. 2003. Preserving Web sites: A data quality approach. In *Proceedings of the 7th International Conference on Information Quality (ICIQ)*.

Cappiello, C., Francalanci, C., Pernici, B., Plebani, P., and Scannapieco, M. 2003b. Data quality assurance in cooperative information systems: a multi-dimension certificate. In *Proceedings of the ICDT International Workshop on Data Quality in Cooperative Information Systems (DQCIS)*.

Catarci, T., and Scannapieco, M. 2002. Data quality under the computer science perspective. *Archivi Computer 2*.

Chapman, A., Richards, H., and Hawken, S. 2006. Data and information quality at the Canadian institute for health information. In *Proceedings of the 11th International Conference on Information Quality*.

Chengalur-Smith, I. N., Ballou, D. P., and Pazer, H. L. 1999. The impact of data quality information on decision making: An exploratory analysis. *IEEE Trans. Knowl. Data Eng. 11*, 6, 853–864.

Corey, D., Cobler, L., Haynes, K., and Walker, R. 1996. Data quality assurance activities in the military health services system. In *Proceedings of the 1st International Conference on Information Quality*. 127–153.

Dasu, T. and Johnson, T. 2003. *Exploratory Data Mining and Data cleaning*. Probability and Statistics series, John Wiley.

Data Warehousing Institute. 2006. Data quality and the bottom line: Achieving business success through a commitment to high quality data. http://www.dw-institute.com/.

De Amicis, F., Barone, D., and Batini, C. 2006. An analytical framework to analyze dependencies among data quality dimensions. In *Proceedings of the 11th International Conference on Information Quality (ICIQ)*. 369–383.

De Amicis, F. and Batini, C. 2004. A methodology for data quality assessment on financial data. *Studies Commun. Sci.* SCKM.

De Michelis, G., Dubois, E., Jarke, M., Matthes, F., Mylopoulos, J., Papazoglou, M., Pohl, K., Schmidt, J., Woo, C., and Yu, E. 1997. Cooperative Information Systems: A Manifesto. In *Cooperative Information Systems: Trends & Directions*, M. Papazoglou and G. Schlageter, Eds. Academic-Press.

De Santis, L., Scannapieco, M., and Catarci, T. 2003. Trusting data quality in cooperative information systems. In *Proceedings of the 11th International Conference on Cooperative Information Systems (CoopIS)*. Catania, Italy.

Dedeke, A. 2005. Building quality into the information supply chain. *Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph*. R. Wang, E. Pierce, S. Madnick, and Fisher C.W., Eds.

DQI. 2004. Data quality initiative framework. Project report. www.wales.nhs.uk/sites/documents/319/DQI_Framwork_Update_Letter_160604.pdf

English, L. 1999. *Improving Data Warehouse and Business Information Quality*. Wiley & Sons.

English, L. 2002. Process management and information quality: how improving information production processes improved information (product) quality. In *Proceedings of the 7th International Conference on Information Quality (ICIQ)*. 206–211.

Eppler, M. and Helfert, M. 2004. A classification and analysis of data quality costs. In *Proceedings of the 9th International Conference on Information Systems (ICIQ)*.

Eppler, M. and Münzenmaier, P. 2002. Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology. In *Proceedings of the 7th International Conference on Information Systems (ICIQ)*.

Falorsi, P., Pallara, S., Pavone, A., Alessandroni, A., Massella, E., and Scannapieco, M. 2003. Improving the quality of toponymic data in the italian public administration. In *Proceedings of the ICDT Workshop on Data Quality in Cooperative Information Systems (DQCIS)*.

Fellegi, I. P., and Holt, D. 1976. A systematic approach to automatic edit and imputation. *J. Amer. Stat. Assoc. 71*, 353, 17–35.

Fisher, C. and Kingma, B. 2001. Criticality of data quality as exemplified in two disasters. *Inform. Manage. 39*, 109–116.

Fraternali, P., Lanzi, P., Matera, M., and Maurino, A. 2004. Model-driven Web usage analysis for the evaluation of Web application quality. *J. Web Eng. 3*, 2, 124–152.

Gackowski, Z. 2006. Redefining information quality: the operations management approach. In *Proceedings of the 11th International Conference on Information Quality (ICIQ)*. 399–419.

Hammer, M. 1990. Reengineering work: Don't automate, obliterate. *Harvard Bus. Rev.* 104–112.

Hammer, M. and Champy, J. 2001. *Reengineering the Corporation: A Manifesto for Business Revolution*, Harper Collins.

Hernandez, M. and Stolfo, S. 1998. Real-world data is dirty: Data cleansing and the merge/purge problem. *J. Data Min. Knowl. Dis. 1*, 2.

Isakowitz, T., Bieber, M., and Vitali, F. 1998. Web information systems - introduction. *Commun. ACM 41*, 7, 78–80.

Isakowitz, T., Stohr, E., and Balasubramanian, P. 1995. RMM: A methodology for structured hypermedia design. *Comm. ACM 58*, 8.

Istat. 2004. Guidelines for the data quality improvement of localization data in public administration (in Italian). www.istat.it

Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P., Eds. 1995. *Fundamentals of Data Warehouses*. Springer Verlag.

Jeusfeld, M., Quix, C., and Jarke, M. 1998. Design and analysis of quality information for data warehouses. In *Proceedings of the 17th International Conference on Conceptual Modeling*.

Kerr, K. and Norris, T. 2004. The development of a healthcare data quality framework and strategy. In *Proceedings of the 9th International Conference on Information Quality*.

Kettinger, W. and Grover, V. 1995. Special section: Toward a theory of business process change management. *J. Manag. Inform. Syst. 12*, 1, 9–30.

Kovac, R. and Weickert, C. 2002. Starting with quality: Using TDQM in a start-up organization. In *Proceedings of the 7th International Conference on Information Quality (ICIQ)*. Boston, 69–78.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. AIMQ: A methodology for information quality assessment. *Inform. Manage. 40*, 2, 133–460.

Lenzerini, M. 2002. Data integration: A theoretical perspective. In *Proceedings of the 21st ACM Symposium on Principles of Database Systems (PODS)*.

Liu, L. and Chi, L. 2002. Evolutionary data quality. In *Proceedings of the 7th International Conference on Information Quality*.

Long, J. and Seko, C. April 2005. A cyclic-hierarchical method for database data-quality evaluation and improvement. In *Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph*, R. Wang, E. Pierce, S. Madnick, and Fisher C.W.

Loshin, D. 2004. *Enterprise Knowledge Management - The Data Quality Approach*. Series in Data Management Systems, Morgan Kaufmann, chapter 4.

Lyman, P. and Varian, H. R. 2003. How much information. http://www.sims.berkeley.edu/how-much-info-2003.

Mecca, G., Atzeni, P., Masci, M., Merialdo, P., and Sindoni, G. 1998. The Araneus Web-based management system. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, L. M. Haas and A. Tiwary, Eds. ACM Press, 544–546.

Mecca, G., Merialdo, P., Atzeni, P., and Crescenzi, V. 1999. The (short) araeneus guide to Web site development. In *Proceedings of the 2nd International Workshop on the Web and Databases (WebDB)* Conjunction with Sigmod.

Motro, A. and Anokhin, P. 2005. Fusionplex: Resolution of data inconsistencies in the data integration of heterogeneous information sources. *Inform. Fusion, 7*, 2, 176–196.

Muthu, S., Withman, L., and Cheraghi, S. H. 1999. Business process re-engineering : a consolidated methodology. In *Proceedings of the 4th annual International Conference on Industrial Engineering Theory, Applications and Practice*.

Nadkarni, P. 2006. Delivering data on time: The assurant health case. In *Proceedings of the 11th International Conference on Information Quality*.

Naumann, F. 2002. Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science, vol. 2261.

NELSON, J., POELS, G., GENERO, M., AND PIATTINI, EDS. 2003. *Proceedings of the 2nd International Workshop on Conceptual Modeling Quality (IWCMQ).* Lecture Notes in Computer Science, vol. 2814, Springer.

OAKLAND, J. 1989. *Total Quality Management*. Springer.

OFFICE OF MANAGEMENT AND BUDGET. 2006. Information quality guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by agencies. http://www.whitehouse.gov/omb/fedreg/reproducible.html.

PERNICI, B. AND SCANNAPIECO, M. 2003. Data quality in Web information systems. *J. Data Semant. 1*, 48–68.

PIPINO, L., LEE, Y., AND WANG, R. 2002. Data quality assessment. *Commun. ACM 45*, 4.

RAGHUNATHAN, S. 1999. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decis. Supp. Syst. 26*, 275–286.

RAHM, E., THOR, A., AUMÜLLER, D., HONG-HAI, D., GOLOVIN, N., AND KIRSTEN, T. June 2005. iFuice information fusion utilizing instance correspondences and peer mappings. In *Proceedings of the 8th International Workshop on the Web and Databases (WebDB).* located with SIGMOD.

RAO, R. 2003. From unstructured data to actionable intelligence. *IT Professional 535*, 6, 29–35.

REDMAN, T. 1996. *Data Quality for the Information Age*. Artech House.

REDMAN, T. 1998. The impact of poor data quality on the typical enteprise. *Comm. ACM 41*, 2, 79–82.

SCANNAPIECO, M., A.VIRGILLITO, MARCHETTI, M., MECELLA, M., AND BALDONI, R. 2004. The DaQuinCIS architecture: a platform for exchanging and improving data quality in Cooperative Information Systems. *Inform. Syst. 29*, 7, 551–582.

SCANNAPIECO, M., PERNICI, B., AND PIERCE, E. 2002. IP-UML: Towards a Methodology for Quality Improvement based on the IP-MAP Framework. In *Proceedings of the 7th International Conference on Information Quality (ICIQ)*. Boston.

SCANNAPIECO, M., PERNICI, B., AND PIERCE, E. 2005. IP-UML: A methodology for quality improvement-based on IP-MAP and UML. In *Information Quality, Advances in Management Information Systems, Information Quality Monograph (AMIS-IQ)*, R. Wang, E. Pierce, S. Madnik, and C. Fisher, Eds.

SESSIONS, V. 2007. Employing the TDQM methodology: An assessment of the SC SOR. In *Proceedings of the 12th International Conference on Information Quality*. 519–537.

SHANKARANARAYAN, G., WANG, R. Y., AND ZIAD, M. 2000. Modeling the manufacture of an information product with IP-MAP. In *Proceedings of the 6th International Conference on Information Quality (ICIQ 2000)*. Boston.

SHANKARANARAYANAN, G. AND WANG, R. 2007. IPMAP: Current state and perspectives. In *Proceedings of the 12th International Conference on Information Quality*.

SHENG, Y. 2003. Exploring the mediating and moderating effects of information quality on firm's endeavour on information systems. In *Proceedings of the 8th International Conference on Information Quality 2003 (ICIQ)*. 344–352.

SHENG, Y. AND MYKYTYN, P. 2002. Information technology investment and firm performance: A perspective of data quality. In *Proceedings of the 7th International Conference on Information Quality (ICIQ)*. DC, 132–141.

STOICA, M., CHAWAT, N., AND SHIN, N. 2003. An investigation of the methodologies of business process reengineering. In *Proceedings of Information Systems Education Conference*.

SU, Y. AND JIN, Z. 2004. A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In *Proceedings of the 9th International Conference on Information Quality (ICIQ)*. 447–465.

US DEPARTMENT OF DEFENSE. 1994. Data administration procedures. DoD rep. 8320.1-M.

VASSILIADIS, P., VAGENA, Z., SKIADOPOULOS, S., KARAYANNIDIS, N., AND SELLIS, T. 2001. ARTKOS: toward the modeling, design, control and execution of ETL processes. *Inform. Syst. 26*, 537–561.

VERMEER, B. 2000. How important is data quality for evaluating the impact of edi on global supply chains. In *Proceedings of the 33rd Haway Conference on Systems Sciences*.

WAND, Y. AND WANG, R. 1996. Anchoring data quality dimensions in ontological foundations. *Comm. ACM 39*, 11.

WANG, R. 1998. A product perspective on total data quality management. *Comm. ACM 41*, 2.

WANG, R. AND STRONG, D. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inform. Syst. 12*, 4.

WORLD WIDE WEB CONSORTIUM. www.w3.org/WAI/. Web accessibility initiative.

ZACHMAN, J. 2006. Zachman institute for framework advancement (ZIFA). www.zifa.com.

Received December 2006; revised December 2007; accepted May 2008