Assessment and analysis of information quality: a multidimensional model and case studies

Laure Berti-Équille*

CEREGE-IRD,
Europôle Méditerranéen de l'Arbois, Av. L. Philibert, BP 80,
13545 Aix-En-Provence, France
E-mail: laure.berti@ird.fr
*Corresponding author

Isabelle Comyn-Wattiau

CEDRIC, Conservatoire National des Arts et Métiers (CNAM)
et ESSEC Business School,
292 Rue Saint Martin, 75141 Paris Cedex 03, Paris, France
E-mail: isabelle.wattiau@cnam.fr

Mireille Cosquer

Service d'Information Médicale,
Institut Curie,
26, Rue des Fossés Saint-Jacques, 75005 Paris
E-mail: mireille.cosquer@curie.net

Zoubida Kedad

Laboratoire PRiSM,
Université de Versailles Saint-Quentin-en-Yvelines,
45 Avenue des Etats-Unis, 78035 Versailles Cedex, France
E-mail: zoubida.kedad@prism.uvsq.fr

Sylvaine Nugier

EDF – R&D,
Département STEP,
6 Quai Watier, BP 49, 78401 Chatou Cedex, France
E-mail: sylvaine.nugier@edf.fr

Verónika Peralta

Antenne Universitaire de Blois,
3 Place Jean Jaurès, 41000 Blois, France
E-mail: veronika.peralta@univ-tours.fr

Samira Si-Saïd Cherfi and
Virginie Thion-Goasdoué

CEDRIC, Conservatoire National des Arts et Métiers (CNAM),
292 Rue Saint Martin, 75141 Paris Cedex 03, Paris, France
E-mail: sisaid@cnam.fr
E-mail: virginie.thion@cnam.fr

Abstract: Information quality is a complex and multidimensional notion. In the context of information system engineering, it is also a transversal notion and to be fully understood, it needs to be evaluated jointly considering the quality of data, the quality of the underlying conceptual data model and the quality of the software system that manages these data. This paper presents a multidimensional model for exploring information in a multidimensional way, which aids in the navigation, filtering, and interpretation of quality measures, and thus in the identification of the most appropriate actions to improve information quality. Two application scenarios are presented to illustrate and validate the multidimensional approach: the first one concerns the quality of customer information at Electricité de France, a French electricity company, and the second concerns the quality of patient records at Institut Curie, a well-known medical institute in France. The instantiation of our multidimensional model in these contexts shows first illustrations of its applicability.

Keywords: information quality; model quality; data quality; quality analysis.

Reference to this paper should be made as follows: Berti-Équille, L., Comyn-Wattiau, I., Cosquer, M., Kedad, Z., Nugier, S., Peralta, V., Cherfi, S.S-S. and Thion-Goasdoué, V. (2011) 'Assessment and analysis of information quality: a multidimensional model and case studies', *Int. J. Information Quality*, Vol. 2, No. 4, pp.300–323.

Biographical notes: Laure Berti-Équille received her MSc in Physics from the University of Paris IX in 1995, her MSc degree and her PhD in Computer Science from the University of Toulon (France) in 1996 and 1999, respectively. From 2000 to 2007, she joined IRISA (INRIA/CNRS/INSA/Univ. of Rennes 1) as a permanent Associate Professor. From 2007 to 2009, she was a Visiting Researcher at AT&T Labs Research, NJ, USA. Since 2011, she is the Director of Research at IRD (Institute of Research for Development) involved in several data quality research projects.

Isabelle Comyn-Wattiau is a Full Professor at the CEDRIC Research Center, at Conservatoire National des Arts et Métiers, French University located in Paris. Her research is mainly related to information system engineering. She has published many papers in international journal and conference proceedings, especially concerning schema integration, ontology for information system design, database reverse engineering, decision system and datawarehouse design, information system evolution and information system quality. She has written and edited several books, mainly about databases. Her main contribution consists of definition of models, methods and approaches for design, development and reengineering of advanced information systems, especially but not only information systems based on datawarehouses and websites.

Mireille Cosquer has been a Statistician since 2001 at the Medical Information Service of the Institut Curie hospital. Being an Engineer CNAM (1997) in data mining specialisation, her interest focused on quality assessment techniques through the electronic patient record to ensure data quality of patient files. She is also in charge of the Identity Vigilance Unit of the hospital since 2008.

Zoubida Kedad is an Associate Professor of Computer Science at University of Versailles St-Quentin en Yvelines (UVSQ). She is a member of the Advanced Modelling of Information Systems (AMIS) Group at PRiSM Laboratory. Her research interests are database design and multisource information systems design. Her works mainly concerns data quality, semantic data integration and evolution. She has been involved in several national and international projects on data integration and data quality.

Sylvaine Nugier works at Electricité De France Research and Developpement, on information retrieval from large amounts of heterogeneous data. The methods and tools used or explored in Sylvaine Nugier's studies, focused on data modelling, textual data analysis, social networks analysis, and information visualisation. In EDF, she has acquired a solid expertise in data quality and participates in various projects to support operational units in the field of sales, marketing and electricity production.

Verónika Peralta is an Assistant Professor at the University of Tours since 2008. She received her PhD in Computer Science from the University of Versailles and the University of the Republic of Uruguay in 2006. Her research interests include quality of data, quality of service, query personalisation, data warehousing and OLAP. She has taught multiple courses since 1996, in several universities in Uruguay, Argentina and France and worked in many research projects in collaboration with Uruguayan, Brasilian and French universities.

Samira Si-saïd Cherfi is an Assistant Professor at the Conservatoire National des Arts et Métiers. She obtained her doctorate degree at the University of Paris 1 – Panthéon Sorbonne. Her research interests include information systems methodologies, information systems quality, methods and tools for information systems engineering, method engineering and quality assessment and improvement. She has co-organised the Quality of Information Systems (QoIS) workshops (QoIS'05, QoIS'06 and QoIS'07) held in conjunction with the ER International Conference. She was involved in projects on quality and security of information systems.

Virginie Thion-Goasdoué obtained her PhD in Computer Science in 2004. Then, she worked as a Research Engineer in the R&D entity of the major integrated energetic utility company in France (Electricité de France). In 2008, she joined the LAMSADE Laboratory of the Univ. Paris Dauphine as an Associate Professor. Since 2010, she is an Associate Professor at the CEDRIC Laboratory, Conservatoire National des Arts et Métiers (CNAM), Paris. Her research interests concern the management of databases and information systems.

1 Introduction

In the past few years, information quality has become a very hot topic both in academic research and in industrial contexts. Researchers attempt to provide formal definitions of information quality enabling measurement and automatic approaches for quality assessment and improvement. Companies use market software tools to compute quality measures from their data, their programmes, their automatic or manual processes to their software and large-scale applications. Metrics and tools computing these measures are plethoric, leading to a huge amount of data and metadata. As a consequence, there is a need for filtering and interpreting this data and metadata, to compare different measures and thus decide the most appropriate actions to improve information quality at various levels, from the quality of data to the quality of the information systems (ISs).

Capitalising on QUADRIS project (Akoka et al., 2007; QUADRIS project, 2009), we argue that it is possible to combine and jointly explore a variety of measures characterising the quality of data and the quality of the data model in order to provide users, designers, and developers with a better understanding of the transversal notion of information quality. In this paper, we propose a multidimensional model gathering all quality measures, obtained from computation on data and models. These measures are defined according to relevant analysis dimensions and stored in a star-like multidimensional database, which eases the navigation, filtering and interpretation of quality measures, and thus the identification of the most appropriate actions to improve information quality. Note that *quality dimensions*, which refers to different facets of information quality (e.g., readability, accuracy or response time), should be distinguished from *analysis dimensions* which refers to analysis criteria (such as assessment date, involved actors or quality goals). Our model is multidimensional from these two complementary points of views.

To validate our approach, we have conducted experiments on different application contexts. Two of these contexts are:

- *Customer relationship management* (CRM) conducted at *Electricité de France* (EDF) (http://www.edf.fr), a French integrated energetic utility company managing all aspects of the electricity business. EDF has a total of 40.2 million customers worldwide (including 28 million in France). A mission of EDF's marketers is to undertake surveys about the energy load curves of their individual customers and ensure the quality of their customers' information stored in their CRM databases.
- Healthcare services and patient management at the *Institut Curie* (http://www.curie.fr). The mission of this French healthcare institute as a public service body since 1921 is treatment and research against cancer. The institute is willing to develop the tools allowing them to meet the French National Health Authority (HAS) requirements, and particularly regarding the quality of patient data management.

In this paper, we describe these application scenarios and illustrate the use of our model and tools to help fulfilling EDF's and Institut Curie's quality goals. The contribution of this paper is twofold:

1. we present a multidimensional data model for the analysis of quality measures
2. we describe two case studies experienced on quality analysis at EDF and at Institut Curie and discuss the benefits and limits of our approach.

The rest of this paper is organised as follows: The second section is devoted to a brief literature review. The third section summarises the QUADRIS project and its unified meta-model for information quality management. The fourth section describes the multidimensional model that we propose for quality analysis. The validation on the applicative scenarios is illustrated in the fifth section. The sixth section presents QBox, the platform implementing the model and services for information quality assessment and analysis. Finally, Section 7 concludes and describes future research work and perspectives.

2 Related work

The main topic of information quality attracted many research works since two decades. It is an interesting theoretical as well as practical domain in which formalisation is more and more needed. Several research projects [e.g., TDQM (Wang et al., 1995) and TQdM (English, 1999)] defined methodologies and experience recommendations for dealing with quality assurance in business ISs. More recently, several projects proposed data quality assessment and improvement techniques in database and data warehousing domains [e.g., DWQ (Jarke et al., 1999), DaQuinCis (de Santis et al., 2003), and Trio (Widom, 2005)]. We point the work of Batini et al. (2009) that proposed a comparison of such methodologies. Moreover, information quality is a crucial problem in companies and organisations, where IS investments must be justified, appreciated, and re-evaluated day after day. In particular, several quality problems have critical impacts in several scientific areas such as environment (Jankowka, 2000; US Environment Protection Agency, 2004) and genetics (Müller and Naumann, 2003; Salanti et al., 2005).

Comprehensive surveys on information quality can be found in Batini and Scannapieco (2006), Berti-Équille (2007) and Peralta (2006). Information quality is generally described through a large set of quality attributes or factors. Literature aims at defining quality factors and metrics (Redman, 1996; Wang and Strong, 1996), proposing quality models including these factors and metrics (Jarke and Vassiliou, 1997; Strong et al., 1997; Moody, 2005; Caballero et al., 2007), enabling the quantitative evaluation of quality factors (Naumann and Rolker, 2000; Pipino et al., 2002) and proposing taxonomies of factors and metrics (Naumann et al., 1999; Wang et al., 1995).

Quality models are mainly hierarchical, thus allowing a structured approach of information quality. Few papers mention non-hierarchical models. Let us mention:

1 The quality cube model, based on three analysis dimensions: users/clients, product/process and efficiency/effectiveness (Rawashdeh and Matalkah, 2006).

2 The star model, containing three significant elements: the procurer, the producer and the product, not structured in a multidimensional form. It is interesting in the way it presents multiple viewpoints (Fitzpatrick, 1996).

3 The multidimensional model for web-based applications quality, based on three analysis dimensions: application domain (e-learning, e-commerce, etc.), lifecycle processes (development, exploitation, maintenance) and quality characteristics (functionality, reliability, usability, etc). This model aims at assessing the quality of applications depending on their respective domains (Malak et al., 2004).

To the best of our knowledge, no contributions have been proposed for modelling the multiple analysis dimensions of quality for decision-making. This notion encompasses both quality of data and quality of the underlying data model. This paper is a step forward in this direction.

Healthcare information quality has been addressed in many research works. In Miettinen and Korhonen (2008), the authors present a case-based analysis of healthcare data quality problems for data about diabetes patients that are combined from different ISs. Campos et al. (2008) present utilities to exploit digital health records and propose a distributed architecture providing tools for the supervision and analysis of healthcare quality, which can be viewed as the quality of its processes. In Coletti (2007), the authors propose a model for evaluating the quality of a healthcare provider focusing on some relevant indicators using conditional probabilities. In van Deursen et al. (2008), a reputation system to determine quality of health personal records is described, using the reputation of the data operator as a quality indication. Kerr and Norris (2008) survey the acquisition and usage of data quality in the delivery and planning of healthcare and discuss the factors that influence data quality. Civan and Pratt (2006) present a multidimensional model for evaluating the quality of health information available on the web, and present an assessment scenario using the MEDLINE (2010) resource. Data quality in this context is considered as a multidimensional construct characterised by four characteristics: content, usage, authorship, and publication quality. The goal of their approach is to characterise and to evaluate the quality of data sources of a specific application domain, while our approach is more generic since it aims at providing a multidimensional model gathering the various dimensions of information quality in an IS.

3 Quality assessment in QUADRIS

Our quality assessment approach is based on the *goal-question-metric* (GQM) paradigm (Basili et al., 1994). Information quality is analysed at three abstraction levels:

1 at the conceptual level, the approach identifies high-level quality goals (e.g., 'reduce the number of returns in customer mails')

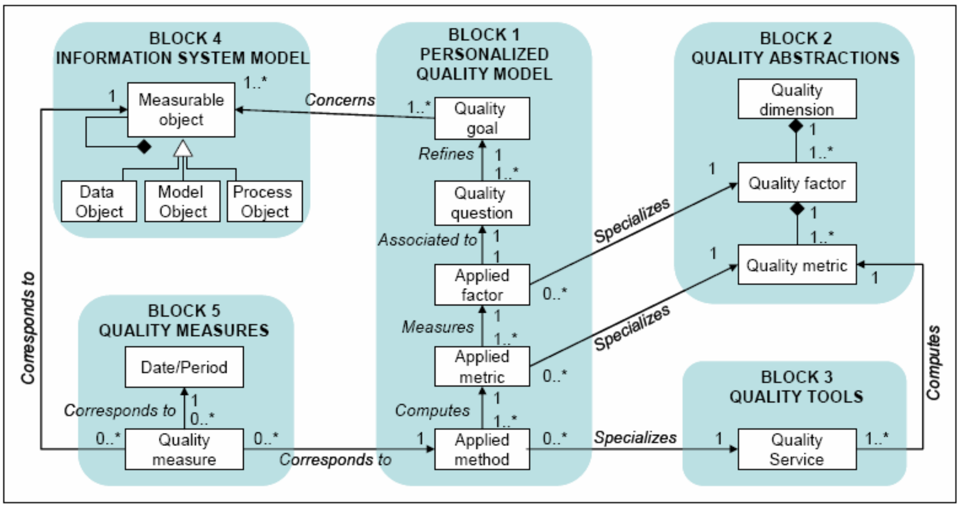
2 at the operational level, it enounces a set of quality questions that characterise the way to assess each goal (e.g., 'which is the amount of syntactic errors in customer addresses?')

3 at the quantitative level, it defines a set of quality measures that quantify the way to answer to each question (e.g., 'the percentage of data satisfying a syntax rule') and a set of measurement methods for computing them.

The core of the approach is a quality assessment meta-model, which allows representing quality concepts and reasoning from them. Figure 1 gives a synthetic view of the meta-model. Case studies in Section 5 will fully illustrate and instantiate it but let us first explain in detail its composition.

The central block (block 1) deals with quality goals following the GQM approach. *Quality goals* represent high-level quality needs, which are refined and decomposed in a set of *quality questions*. The answer to a quality question is defined by choosing and refining a *quality factor* which best characterises the question, a set of *quality metrics* which are appropriate to measure this factor and a set of *methods* of measurement of this metric. A method associated to a metric corresponds to a specific algorithm used for the computation of the quality measure for this metric. Quality factors and metrics are chosen from a library of generic quality concepts (block 2 of the meta-model); measurement methods are chosen from a library of available quality services (block 3 of the meta-model) and bound to the corresponding IS objects (block 4 of the meta-model).

Figure 1 Quality assessment meta-model (see online version for colours)



The second block constitutes an extensible library of abstract data types which will be used to characterise specific quality goals. The main abstractions are: *quality dimensions* which capture a high-level facet of information quality, *quality factors* which represent particular aspects of quality dimensions and *quality metrics* which are the instruments used to measure quality factors. The third block constitutes a library of measurement methods. It is decoupled from the second block in order to manage a large collection of external tools, listed in a service registry.

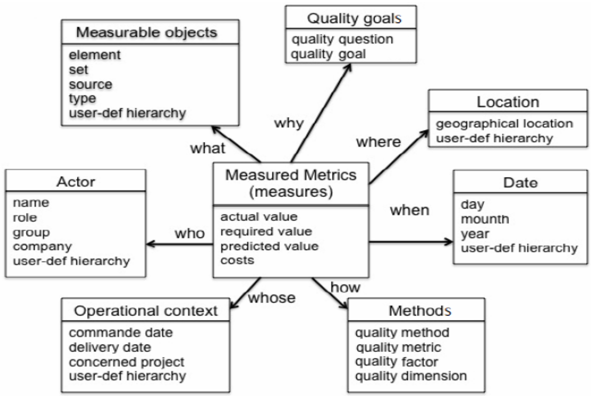
The fourth block refers to the IS model and to the processes which operate on the instances of this model. Each object type, being either data, a model or a process, is called a measurable object if it is subject to a qualitative evaluation within a quality goal.

The fifth block deals with quality measurements which are necessary for evaluating quality questions and diagnosing information quality. Quality measures represent the result of executing a measurement method (for evaluating a quality goal), for a measurable object, at a given instant or during a period of time. Results of successive quality measurements serve to analyse behaviours and trends of the measured objects. Generally, improvement actions are taken based on this analysis. A detailed description of the meta-model can be found in Etcheverry et al. (2008).

4 Multidimensional analysis of quality measures

Quality measures (block 5 of the meta-model in Figure 1) are stored in a star-like database schema which facilitates their aggregation, the computation of complex indicators and the analysis of correlations among the measures. This section describes the multidimensional data model and presents the spectrum of analysis techniques provided by this model.

Figure 2 Multidimensional data model for analysis of quality measures



4.1 Multidimensional data model

The quality meta-model presented in Figure 1 explicitly shows that each quality measure is associated to a date (or period), a measurable object of the IS and an applied method. The latter determines quality metrics, factors and dimensions, as well as quality questions and goals. In addition, a context can be derived from IS objects (e.g., the geographical location of the IS) and user information can be obtained from the goals (e.g., who defined each goal). Figure 2 presents a star schema directly derived from these relationships. The schema contains as analysis criteria:

- *Date:* Indicates *when* quality measures were taken. The dimension includes the classical day-month-year hierarchy as well as additional user-defined periods.
- *Measurable objects:* Indicates *what* is measured, i.e., which objects are examined for computing their quality. The main hierarchy consists of element (e.g., cells in a table or entities in a model), set (e.g., tables, packages) and source (e.g., database, application). A secondary hierarchy indicates the type of object (data, model or process). Other hierarchies can be introduced for grouping objects according to domain-specific relationships (for example, in biomedical applications, the laboratory that produced data is usually a drill-up criterion).
- *Quality methods:* Indicates *how* quality measures were taken. The dimension hierarchy corresponds to blocks 2 and 3 of the meta-model (quality service, quality metric, quality factor and quality dimension in Figure 1).
- *Quality goals:* Indicates *why* these measures were taken, i.e., the purposes of quality analysis. The dimension hierarchy corresponds to block 1 of the meta-model (quality question, quality goal).
- *Location:* Indicates the *geographical location* to which measures are associated to. This context is generally deduced from measurement objects, for example, a datum representing the electricity consumption of a house can be associated to the geographical location of the house. This hierarchy is domain-dependent.

- *Actor:* Indicates *who* conducted quality measurement (the persons that defined quality goals, chose quality methods…). Typical hierarchies include group and enterprise, but may be personalised for a specific application.
- *Operational context:* Quality goals come from a business problem or business goal. Such a problem is linked to an operational context described by a request date, deliverable date, sponsor, operational constraints, etc.

The multidimensional schema contains as indicators:

- *Actual quality value:* Refers to the quality measures that are actually computed by measurement methods.
- *Required quality value:* Refers to the quality bounds that are tolerated by users. These bounds are usually indicated when expressing and refining a quality goal. Actual values are said to be good if they do not overflow required values.
- *Predicted quality value:* Refers to the quality values that users expect to obtain or the values estimated by other profiling tools. They are generally compared to actual values in order to reassert or contradict a hypothesis about data quality.
- *Non-quality cost:* Refers to the cost (e.g., money, time, human resources) caused by poor quality objects and assumed by the company. Cost estimation is application-dependent. It may be defined when expressing and refining quality goals. It may include non-quality cost, quality improvement cost, or quality measurement cost (English, 1999).

For implementing the multidimensional data model, three major problems have to be considered:

- *Additivity:* Quality values can be averaged, but in most cases, domain-specific roll-ups are desired. For example, when totalising values corresponding to different quality metrics, in order to obtain an aggregated value for a quality factor, different types of weights can be applied. Analogously, a particular quality question may indicate how to aggregate individual quality values in order to answer the question. These specialised roll-up operations may be different for each type of IS object, quality factor, quality goal, operational context and actor. This forces roll-ups to be computed on query time.
- *Dynamic dimensions:* As previously discussed, many dimensions should be analysed by user-defined hierarchies, which may be of various complexity and size. An implementation of the model must support the management of dynamic dimensions.
- *Amount of data:* As several quality measures may be taken for individual data elements (e.g., each cell of a table), storage constraints have to be taken into account.

4.2 Analysis of quality measures

The rationale of the multidimensional quality model is threefold. First, it has been designed for the purpose of exploring quality measures with various analytic tasks. Second, it can be used for scoring data with respect to user-defined or application-driven quality requirements and prioritising tasks for quality improvement. Finally, it can be used for quality prediction and forecasting:

1. *Quality exploration* includes three tasks described as follows:
   - *Quality diagnosis:* once the quality measures are computed, they instantiate the multidimensional model and can be browsed in order to diagnose, visualise, and understand the quality of the information, both at the instance (data quality) and schema levels (model quality).
   - *Metric selection:* based on the quality measures provided by the multidimensional model, this task allows the user to experimentally compare a variety of metrics in order to choose the most appropriate ones for highlighting a suspected or known phenomenon.
   - *Metric correlation:* this task analyses the relationships between the various stored quality measures, looking for dependencies or correlation.
2. *Quality scoring* includes two tasks described as follows:
   - *Data recommendation:* based on the highest quality measures computed from the data managed by the IS, this task provides and associates quality guarantees to the data queried by the users.
   - *Task recommendation:* based on the lowest quality measures computed from the data and the data model, this task provides priorities for scheduling cleaning and corrective actions to improve overall information quality.
3. *Quality prediction* aims at computing the trends and forecasting information quality over time based on the history of quality measures and various input prediction models.

From a more technical point of view, the multidimensional data model is a natural entry for OLAP (e.g., Oracle Express or Olap option, Microsoft Analysis Services, IBM TM1) and statistical (e.g., SAS or R) tools.

5 Case studies

In this section, we illustrate the use of our quality assessment framework in two different real contexts. The model was discussed and improved through the dialogue with companies which need quality frameworks to check their IS qualities.

5.1 A CRM scenario at Electricité de France

The *EDF* group is an integrated energetic utility company managing all aspects of the electricity business. In this paper, we focus on a commercial aspect of the group. EDF has a strong commercial footing in Europe, with a total of 40.2 million customers worldwide (including 28 million in France). In this section we present the instantiation of the quality meta-model presented above for a CRM application at EDF.

5.1.1 Quality problems

For this scenario, we consider a business user, a marketer, who has to undertake a survey about the energy load curves of EDF customers for a targeted year, e.g., 2007. Customers' information is stored in a CRM database supporting the management of major and small business French markets. This database results from the integration of several heterogeneous operational databases (e.g., front office databases) and some external databases (e.g., geographical referential databases for postal addresses). Data quality problems in such a database are due either to a poor data quality in various operational sources (e.g., if a meter breakdown then the resulting load curve can be incomplete, if a commercial agent misunderstands the postal address of a customer then the address can be inaccurate, etc.) or to integration difficulties (e.g., problems in schema mapping or data reconciliation stages).

In practice, the preliminary task of a marketing survey is to select the studied population. Thus the marketer first and foremost needs to characterise the quality of accessed (or retrieved) information. As an example, the marketer will *select the largest possible set of contacts ensuring that the number of wrong phone calls is minimal*. This is the *operational goal*. The customers having an active contract in 2007 are first identified and checked to make sure that they are individuals (not companies). For each of these customers, the information about the corresponding energy load curves is considered and only the customers with a complete history of load curves are kept for further analysis. The information about the clients is then controlled (e.g., phone numbers or customer price code), and the consistency between the invoices corresponding to each customer and the records of consumed energy are checked. The size of the resulting set has to be statistically significant in order to be a meaningful basis for the survey. The sponsor of this quality survey is the marketing service. To instantiate our model (block 1 of Figure 1), we define the operational goal: *producing the largest possible set of customers that minimises wrong calls* and the following quality goals:

- G1: improve customers contact information for marketing requirements needs
- G2: improve the quality of energy load curves.

5.1.2 Identification of quality metrics

Six quality questions are defined in order to refine the quality goals defined above; they are listed in Table 1. The first question deals with the need of distinguishing individual customers from companies; this information is not always available in all operational sources and often leads to wrong classifications of customers. Questions Q1.2 and Q1.3 are concerned with the validity of customers' information. Question Q1.4 aims at quantifying the portion of customers that are taken into account. Question Q2.1 deals with the availability of data for computing the history of customers' energy load curves for a given period. The last question aims at verifying the coherence of energy load information. The *quality goals* dimension is then instantiated with these values.

Table 1 Instantiation of *quality goals* for the CRM scenario at EDF

Goal	*Question*	
G1	Q1.1	Have customers an ongoing contract? Are they individuals or companies?
	Q1.2	Are customers' phone numbers valid?
	Q1.3	Are customers' contracts valid?
	Q1.4	Are all the individual customers present in the resulting set?
G2	Q2.1	Which are the clients with complete recorded history of consumed loads of energy?
	Q2.2	Are the invoice and the consumed load of energy consistent?

Quality questions are declined in terms of quality dimensions and quality factors of our meta-model. A set of eight quality metrics were defined for answering to quality questions, and a set of measurement methods are used for assessing these metrics, as illustrated in Table 2. The corresponding quality factors are defined in Table 3.

Table 2 Instantiation of *quality methods* for the CRM scenario at EDF

Q#	*Factor*	*Metric*		*Method*
Q1.1	Semantic correctness (accuracy)	M1	Ratio of individuals among the customers	*CheckReferential:* Compare customer information with companies' directories
		M2	Ratio of records that are unlikely to be individuals	*DictionaryLookUp:* Check that the denomination does not contain usual status (Ms, Mrs…) but legal enterprise status (Group, Holding, Corp….) in a dictionary of denominations (SQL procedure enclosed in an ETL workflow)
Q1.2	Syntactic correctness (accuracy)	M3	Ratio of phone numbers having the required format	*Aggregation:* Method provided by the DataFlux (http://www.dataflux.com/) tool
Q1.3	Syntactic correctness (accuracy)	M4	Ratio of customers with a valid tariff code in their contract	*DictionaryLookUp:* Comparison of tariff code to the content of a tariff dictionary
Q1.4	Coverage (completeness)	M5	Difference between the expected number of customers and the size of the resulting set	*Aggregation:* Count of the total of customers and evaluation of the relevance of this value (human validation)
Q2.1	Density (completeness)	M6	Ratio of NULL energy load values for each customer's record	*CheckNull:* Check the presence of NULL values in load curves (SQL queries)
	Coverage (completeness)	M7	Number of records of energy load for each customer	*Count:* SQL query on the database
Q2.2	Record integrity (consistency)	QM8	Ratio of customers for which the difference between invoice and consumed energy load exceeds a threshold	*CheckRule:* SQL queries on the database

Table 3 Quality factors for both, the CRM and medical scenarios

Factor	*Description*
Coverage	Describes whether all required entities are present in the IS (Naumann et al., 2003)
Density	Describes whether all data values are present (not null) for required attributes (Naumann et al., 2003)
Semantic correctness	Describes how well data represent states of the real-world (Wang and Strong, 1996)
Syntactic correctness	Expresses the degree to which data is free of syntactic errors such as misspellings and format discordances (Müller and Naumann, 2003)
Record integrity	Expresses the degree to which data satisfies a set of inter-attribute integrity constraints (Rahm and Do, 2000).
Response time	Expresses the time elapsed between an event and its response (Oasis, 2008)

5.1.3 Instantiation of the multidimensional model

The *measurable objects* dimension (given in Figure 2) in our scenario follows the classical element-set-source hierarchy, where *element* represents a table cell (a value of a record describing a customer), *set* represents a table attribute and *source* represents the source database where data was extracted from. Several operational and external databases are used as sources. The accessed tables and attributes depend on the quality metrics to be computed, for example, for quality metric M2, we access to two attributes of the *customers* table of a given source (namely *civility* and *name*). A user-defined hierarchy enables aggregating elements by customer and type of customer.

The *location* dimension is instantiated with a user-defined hierarchy, consisting of France geographical locations and EDF-defined zones.

Three major actors are involved in this survey: a sponsor (from the EDF marketing service), a data quality expert (from the EDF R&D entity) and an external performer company (A.I.D., http://www.aid.fr). We instantiate the *actor* dimension of Figure 2 with these values as illustrated in Table 4. These actors are common to all quality metrics.

Table 4 Instantiation of *actors* for the CRM scenario at EDF

Name	*Role*	*Group*	*Company*
Anonymous	Sponsor	Marketing entity	EDF
S. Nugier	DQ Expert	R&D entity	EDF
B. Laboisse	Performer		A.I.D.

The *operational contexts* dimension in Figure 2 is instantiated with information about the concerned quality survey (sponsored by the marketing service). Finally, the *date* dimension is instantiated with all dates in the analysed period (2007).

The crossing of previous dimensions corresponds to a set of facts that are stored in the fact table. We register four measures: actual quality value, required quality value, predicted quality value and, when possible, the non-quality cost. A non-quality cost is usually difficult to measure in terms of monetary and human costs but it can be expressed in terms of custom indicators, for example, the number of NPAI[1] (return to sender) for incorrect postal addresses.

The *date* dimension allows us to store these values with a timestamp and thus to follow their evolution over time. This is a good way to detect impacts of improvement actions or, more generally, of any modification of the IS (e.g., integration of a new source, modification of the conceptual data model, etc.).

In addition, the *source* attribute of the *measurable objects* dimension allows us to compare quality measures per data source. It can help to improve the CRM database feeding process either by choosing most reliable sources or detecting data feed problems.

5.1.4 Analysis of quality measures

Among all analyses that can be made, EDF is especially interested in being able to capitalise information quality diagnosis process and results, and perform IS improvement.

The multidimensional model provides an efficient natural way to store results of data quality measures and see values by dimensions of analysis at different levels of aggregation. For exploitation, data is well-suited to be accessed by reporting, data quality or OLAP tools. Analysis of data quality measures can lead to recommendation for IS improvement. For example, if phone numbers are invalid in a specific table, this could mean that the data feeds process or the database schema has to be re-examined.

5.2 Healthcare scenario at Institut Curie

The Institut Curie is a healthcare institute specialised in treatment and research against cancer. Following the nation-wide initiative of the French National Authority for Health (Haute Autorité de Santé, HAS, available at http://www.has-sante.fr/) for the improvement of information quality in the public French healthcare system, the Institut Curie is developing quality procedures to meet the HAS requirements, and particularly regarding the quality of data management and patient records (COMPAQH project, 2008). In the following, we describe the usage of our quality assessment techniques in the Institut Curie's scenario in order to diagnose quality of patient files, calculate quality indicators demanded by the HAS and help in the identification of quality improvement actions.

5.2.1 Quality problems

One key issue for the Institut Curie in order to meet the quality requirements of the HAS is to assess the quality of patient management. In this context, the focus is mainly *on the compliance of the medical files and the quality of the service provided to the patient*. The former can be roughly viewed as an aggregation of the compliance of the patient file's elements (adequate format, presence of all relevant documents, traceability of medical acts, etc.), while the latter concerns the relation with the patient (traceability of the evaluation of pain during the patient's stay, the ability to detect troubles related to a given pathology, the time elapsed between the end of a hospitalisation journey and the shipping of the corresponding letter to the patient and their local doctor, etc.). In order to illustrate our approach, we will focus on two aspects of the relation with the patients: the content of the patient file on the one side, and the time needed to ship the final letter on the other side. When this time delay exceeds eight days, both specific IS-centred corrective actions and administrative service reorganisation are needed.

For the considered operational goal, namely, *the compliance of the patient files and the quality of the service provided to the patient*, several quality goals were defined. We analyse hereafter two representative quality goals, defined as a first step in order to improve Institut Curie's information quality:

- G1: keep patient's records complete for both medical and administrative needs
- G2: reduce the delays in sending the *end of stays* letters.

The patient file is the heart of the Institut Curie's IS, which consists of several, interconnected, source applications. The patient file is composed of three main components: the medical file, the healthcare file and the administrative file, each one containing multiple documents in different formats (reports, letters, images, prescriptions, etc.). More than a thousand new documents are generated per day at the Institut Curie. In addition, some metadata describing the documents (e.g., NIP – *identification permanent number*, type of medical act, medical act date, etc.) are stored in a relational database and serve as index for document search. The set of metadata that is essential for quality assessment and improvement was defined to take into account not only the needs for patient information management but also for other sectors of the institute (French medical information regulation, clinical research, piloting, invoicing, continuous cancer investigation, bio-statistics, etc.) (Civan and Pratt, 2006). This data includes fixed data (e.g., sex, date of birth) but also evolving data (e.g., last contact date). Hence, quality evaluation lies in the satisfaction of quality rules for both the patient file documents and the metadata about those documents.

5.2.2 Identification of quality metrics

The first quality goal concerns the contents of the patient file. Two aspects are important:

1. the file should contain all required documents
2. these documents should be well formatted and should correspond to the appropriate patient.

Several problems are caused by the inadequate format of documents and their metadata. This is worsened by the important number of missing values in document metadata. Consequently, the quality goal was refined in several quality questions (listed in Table 5) concerning the completeness and accuracy of the medical files, i.e., in what extent the file contains all relevant documents and metadata and in what extent documents and metadata correspond to real-world and well-formatted information about the patient.

The second quality goal concerns the indicator of the delay between the date when the patient leaves the hospital and the shipping date of the *end of the stay* letter. For calculating this business indicator, a significant number of patients has to be analysed (to ensure acceptable statistical power) with taking into account the impact of missing values in the patient files. Our methodology for assessing this goal is decomposed as follows:

1. selection of the hospitalisation journeys with a duration greater than 24 hours
2. random sampling of the stratified population of patients depending on the type of healthcare services they have received during their hospitalisation (e.g., obstetric surgery, generalist care)
3. calculation of management indicators.

Hence, quality questions (listed in Table 5) concern the completeness of the documents of the targeted patients and the response time of the shipping processes.

Table 5 Instantiation of the quality assessment meta-model in medical case at Institut Curie

Goal		*Questions*	*Factors*		*Metrics*	*Methods*
G1	Q1.1	Is the patient file complete? Has it all the required documents?	Coverage	M1	Ratio of required documents that are present in the patient file. The required documents are determined according to the hospitalisation protocol for the considered pathology	*FileSearch:* Check the presence of the documents according to the patient's pathology protocol (file lookup)
				M2	Ratio of documents (indexed in the metadata database) that are present in the patient file	*FileSearch:* Check the presence of the documents according to database metadata
	Q1.2	Are the documents in the patient file correctly indexed (document metadata)?	Semantic correctness	M3	Ratio of document metadata correctly indexed	Manual validation of a sample of documents (human processing)
	Q1.3	Have the documents in the patient file all the required fields?	Density	M4	Ratio of mandatory metadata that are not null	*CheckNull:* Check the presence of mandatory metadata (SQL queries)
	Q1.4	Do the documents in the patient file correspond to the patient's medical acts?	Semantic correctness	M5	Ratio of documents effectively belonging to the patient	*CheckReferential:* Cross check between documents metadata and the medical acts databases (SQL procedures)
	Q1.5	Are the documents in the patient file in the expected format? Are they readable?	Syntactic correctness	M6	Ratio of readable files	*OpenDoc:* Check for the documents load with the appropriate software (scripts)
				M7	Ratio of well-formatted document metadata	*CheckRule:* Check for format constraints in the champs of documents metadata (SQL procedures)

Table 5 Instantiation of the quality assessment meta-model in medical case at Institut Curie (continued)

Goal	*Questions*		*Factors*	*Metrics*		*Methods*
G2	Q2.1	Is there a letter for the end of the patient hospitalisation?	Density	M8	Patients with exit letter	*FileSearch:* Check the presence of the exit letter
	Q2.2	Which letters cannot be considered?	Density	M9	Hospitalisation journeys with follow-up doctor	*CheckNull:* Check the presence of metadata
	Q2.3	Which hospitalisation journeys have no date stamps?	Density	M10	Patients without exit date	*LNnull:* Check the presence of metadata (query on the Lotus Notes database)
	Q2.4	Are exit letters shipped within an eight days delay?	Response time	M11	Ratio of patients that received the exit letter within a eight days delay	*LNAggregation:* Count patients where the difference between patient exit and shipping of exit letter is lower than eight days (SQL procedure)
				M12	Average time for shipping exit letters	*LNAggregation:* Average difference between patient exit and shipping of exit letter
				M13	Maximum time for shipping exit letters	*LNAggregation:* Maximum difference between patient exit and shipping of exit letter

Quality questions were associated to quality factors (as described in Table 5) and several metrics were designed for measuring them. Note that some questions were declined in several metrics in order to provide different indicators or to represent various views. Finally, a set of measurement methods was defined for computing these metrics. Some methods (namely *CheckNull*, *CheckRule*, *CheckReferential*, etc.) were reused from the standard catalogue of QBox, the toolbox for data quality assessment developed in the QUADRIS project (Etcheverry et al., 2008) as we will describe in Section 6. Other methods were explicitly implemented for this scenario either by writing simple scripts (i.e., *FileSearch*, *OpenFile*) or by adapting QBox methods to specific platforms (i.e., *LNnull*, *LNaggregation*). A human processing method was also been carried out. Note that a same method can be used for computing several metrics, by defining the appropriate parameters (e.g., format rule, referential). In addition, a method may be used several times (on different objects or with different parameters) in order to compute or update the value of a metric.

5.2.3 Instantiation of the multidimensional data model

In order to instantiate the multidimensional model proposed in Section 3 with our medical scenario, five dimensions are considered: quality goals, operational contexts, dates, quality methods and measurable objects.

The quality goals in our scenario are those presented above (G1 and G2) with the goal-question-metric hierarchy presented in Table 5. The *operational contexts* dimension corresponds to a user-defined hierarchy representing the type of healthcare service (surgery, obstetric etc.). The *date* dimension indicates the date of the measurements, which are not periodic but are triggered by quality analysts according to their measurement plans. The *quality methods* dimension is instantiated with three quality dimensions: completeness (Naumann et al., 2003), accuracy (Peralta, 2006) and performance (Oasis, 2008), the two former concerning data quality and the latter concerning process/service quality. Quality factors are included in Table 3. Completeness factors (coverage and density), accuracy factors (semantic correctness, syntactic correctness) and performance factor (response time), as well as quality metrics and methods are listed in Table 5. The *measurable objects* dimension in our scenario follows a user-defined hierarchy: service-patient-document, where *service* represents a classification of patients according to the main involved healthcare service, *patient* represents a person (and their patient file) and *document* represent a document of the patient file.

Regarding the measures of the multidimensional model, only actual quality values and predicted quality values are considered at the moment for our medical scenario. The former are computed by executing measurement methods and storing their results; the latter are taken from previous estimations resulting from early projects at the Institut Curie.

5.2.4 Analysis of quality measures

The quality analysis model experienced at the Institut Curie can be used in different ways to achieve the quality goals. The first type of analysis concerns the diagnosis of the quality of patient files. The results are important for improving the overall quality indicators, as most HAS indicators depend on the quality of patient files. The multidimensional analysis of the quality metrics M1 to M7 allows determining the common features of the patient files that are more incomplete and/or inaccurate. In particular, the *operational context* dimension (representing healthcare services) allows targeting adequate policies for specific services.

In addition, several action plans were defined by the quality assurance steering committee (responsible for the quality of patient files) at the Institut Curie:

1 identification (and fusion) of duplicate files

2 implementation of input controls in the document database

3 storage of relevant metadata about documents

4 capture of alerts on atypical data

5 capture of undesirable events.

The analysis of measures at different periods allows quantifying the effect of such policies on data quality. The outcomes of action plans are used for the proposal of new corrective actions in a six-month basis.

The second type of analysis concerns the calculation of quality indicators to be provided to the HAS. We illustrate the procedure followed to calculate the delay between the patient exit and the shipping of the exit letter. A series of quality measures is analysed in order to determine a sample of patient files and calculate the indicator. A first sample is randomly selected from the hospitalisation journeys with duration greater than 24 hours. Then, quality metrics M8, M9 and M10 allow filtering the patients that do not satisfy these quality criteria. A second random selection is performed among the remaining patients, obtaining a sample of 80 patient files. Quality metric M12 computes the DEC indicator for the obtained sample. Quality metrics M11 and M13 calculate associate measures for further analysis.

The first measurement was (manually) executed in the second semester of 2006, and recalculated on a six-month basis. This procedure is currently executed at different dates along each semester (e.g., after implementing new input controls or new improvement policies), allowing to obtain different snapshots of the indicators and follow their evolution.

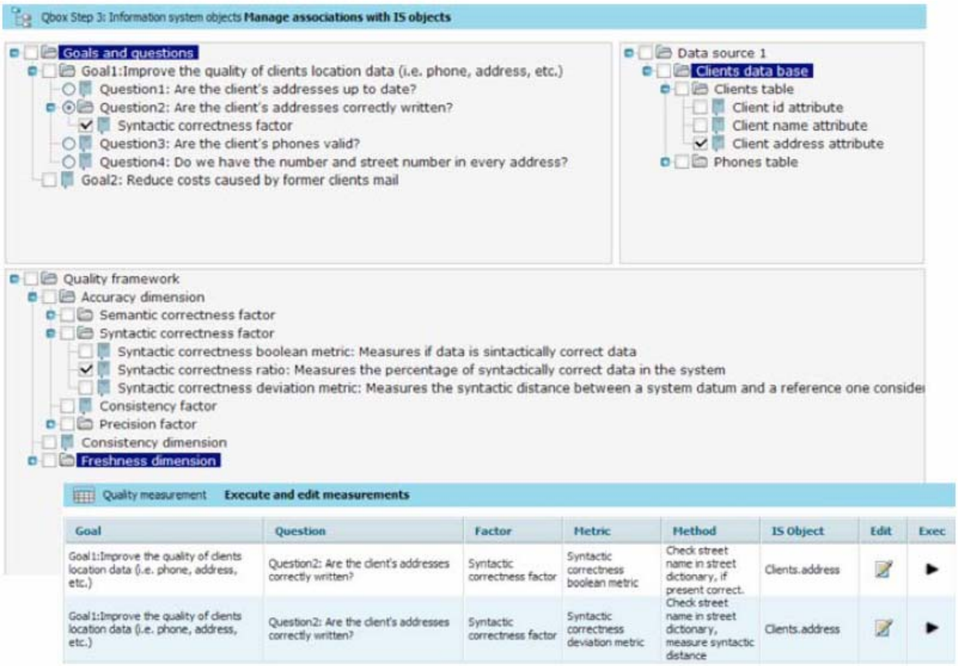
6 QBox: design and implementation

We developed a platform, named QBox to validate our approach on various operational scenarios for the QUADRIS project (Etcheverry et al., 2008). QBox was implemented as a Java web application, with user interfaces for managing the different entities of the multidimensional model and executing measurement methods to assess information quality in our application scenarios. The main functionalities of QBox include:

- Management of an extensible library of quality dimensions, factors, metrics and measurement methods. Methods for retrieving and editing quality concepts and incorporating new ones can be specified and invoked as web services. We have chosen a tree-like structure to show this information to the user. An interface allows the user to define new methods as new services or methods that invoke external routines.

- Definition and storage of user's quality goals and questions. We provide methods for defining and editing quality goals and decomposing them into quality questions. A drag-and-drop interface allows browsing among IS objects and associating them with the relevant questions. This association allows tracking the measures of IS objects quality with respect to specific questions. Analogously, quality factors can be instantiated and associated to questions in a drag-and-drop way. This interface is the starting point for configuring a new quality-assessment application in QBox.
- Association of quality metrics and measurement methods with quality questions. Here, the quality analyst determines what is going to be measured. To this end, a drag-and-drop interface facilitates the browsing among the library of quality concepts (factors, metrics and methods) and their parameterisation. New metrics and methods can be easily defined, either by modifying existing ones or by defining them from scratch.
- Execution of measurement methods for individual IS objects (or all objects) involved in a given quality goal, and persistency management of the obtained quality measures. Specifically, QBox keeps logs of quality measures.
- Show results, allowing the visualisation of trends and correlations. Quality values are stored in a multidimensional way, which allows the comparison of different assessment strategies, the discovery of trends and the exploration of interdependencies among quality factors. The storage of historical values also allows exploring which measurement methods are best suited for each situation and managing quality evolution.

Figure 3 QBox screenshot: instantiation of the GQM approach and quality measurement (see online version for colours)



Qbox Step 3: Information system objects **Manage associations with IS objects**

□ Goals and questions
□ Goal1:Improve the quality of clients location data (i.e. phone, address, etc.)
○ Question1: Are the client's addresses up to date?
● Question2: Are the client's addresses correctly written?
☑ Syntactic correctness factor
○ Question3: Are the client's phones valid?
○ Question4: Do we have the number and street number in every address?
□ Goal2: Reduce costs caused by former clients mail

□ Data source 1
□ Clients data base
□ Clients table
□ Client id attribute
□ Client name attribute
☑ Client address attribute
□ Phones table

□ Quality framework
□ Accuracy dimension
□ Semantic correctness factor
□ Syntactic correctness factor
□ Syntactic correctness boolean metric: Measures if data is sintactically correct data
☑ Syntactic correctness ratio: Measures the percentage of syntactically correct data in the system
□ Syntactic correctness deviation metric: Measures the syntactic distance between a system datum and a reference one conside
□ Consistency factor
□ Precision factor
□ Consistency dimension
□ Freshness dimension

Quality measurement **Execute and edit measurements**

Goal	Question	Factor	Metric	Method	IS Object	Edit	Exec
Goal1:Improve the quality of clients location data (i.e. phone, address, etc.)	Question2: Are the client's addresses correctly written?	Syntactic correctness factor	Syntactic correctness boolean metric	Check street name in street dictionary, if present correct.	Clients.address	✎	▶
Goal1:Improve the quality of clients location data (i.e. phone, address, etc.)	Question2: Are the client's addresses correctly written?	Syntactic correctness factor	Syntactic correctness deviation metric	Check street name in street dictionary, measure syntactic distance	Clients.address	✎	▶

Figure 3 shows how the QBox has been used for the EDF's case study, in particular for instantiating the GQM approach on client addresses and contact information and the measurement of the syntactic correctness factor on clients' addresses with two metrics (Boolean and deviation).

The implementation of QBox is based on Google Web Toolkit technology (GWT, http://code.google.com/intl/fr/webtoolkit/) version 1.7.1 and the extension LGPL SmartGWT (http://code.google.com/p/smartgwt/) version 1.3. Deployment was carried out with a JBoss 3 container, PostgreSQL and MySQL DBMSs. The data access layer encapsulates the access to IS objects and implements persistence mechanisms via Hibernate 3 over the QBox Repository stored in a PostgreSQL database. The Logic layer contains the implementation of the measurement methods and the analysis component. The presentation layer is implemented GWT 1.7.1 and SmartGWT 1.3 component in order to show the measurement results.

7 Conclusions

In this paper, a novel approach for assessing and exploring the multifaceted notion of information quality has been proposed. Our main contribution is a multidimensional model that can capture a large variety of measures for characterising the quality of data and the quality of the underlying data model and data processes. Its goal is to provide users, designers, and developers with a better understanding of the transversal notion of information quality. This model facilitates the navigation, filtering and interpretation of quality measures, and thus the identification of the most appropriate actions to improve information quality. Two real-world case studies have been described to illustrate the applicability of our approach. This approach has three major outcomes. First, it allows exploring relationships among quality concepts (quality metrics, quality factors, quality dimensions, etc.). Secondly, it considers user's preferences on quality requirements. To design effectively operational IS, the IS engineering process has to consider the multiple facets of quality: from the quality of the data, the quality of the underlying data model to the quality of the software and applications. Our multidimensional model is a first step for measuring and exploring the complex notion of quality in a holistic way.

The multidimensional model has been instantiated for EDF's operational scenario and also in the medical context of Institut Curie showing first evidences of its applicability. The case studies allowed a first validation of the approach. The implementation of automated techniques for quality assessment allowed the cross-analysis of quality measures and the follow-up of their evolution. New quality measures may be defined after plan outcomes, enriching the organisations' quality model.

The next step of our work will consist in intensively populating the model with measures and conduct specific statistical analyses to detect dependencies and trends between quality factors and manage quality goals evolution.

Acknowledgements

The QUADRIS project has been partly funded by the ANR ARA Masses de Données Program (2005-2008), 'Grant ARA-05MMSA-0015'.

References

Akoka, J., Berti-Equille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué-Thion, V., Kedad, Z., Nugier, S., Peralta, V. and Sisaid-Cherfi, S. (2007) 'A framework for quality evaluation in data integration systems', *Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS'2007)*.

Basili, V., Caldiera, G. and Rombach, H.D. (1994) 'The goal question metric approach', *Encyclopedia of Software Engineering*, pp.528–532, John Wiley & Sons, Inc.

Batini, C. and Scannapieco, M. (2006) *Data Quality: Concepts, Methodologies, Techniques*, Springer-Verlag.

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009) 'Methodologies for data quality assessment and improvement', *Computing Surveys (CSUR)*, Vol. 41, No. 3, pp.1–52.

Berti-Équille, L. (2007) *Quality Awareness for Data Managing and Mining*, Juin 2007, Habilitation à Diriger des Recherches, Université de Rennes 1, available at http://www.irisa.fr/Laure.Berti-Equille/Habilitation-Laure-Berti-Equille.pdf.

Caballero, I., Verbo, E.M., Calero, C. and Piattini, M. (2007) 'A data quality measurement information model based on ISO/IEC 15939', *Proceedings of the 12th ICIQ*, MIT, Cambridge, MA.

Campos, M., Morales, A., Juárez, J., Sarlort, J., Palma, J. and Marín, R. (2008) 'Intensive care unit platform for health care quality and intelligent systems support', *Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence (DCAI'08)*, pp.366–374.

Civan, A. and Pratt, W. (2006) 'Supporting consumers by characterizing the quality of online health information: a multidimensional framework', *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*.

Coletti, G., Paulon, L., Scozzafava, R. and Vantaggi, B. (2007) 'Measuring the quality of health-care services: a likelihood-based fuzzy modeling approach', *Proceedings 9th European Conference of the Symbolic and Quantitative Approaches to Reasoning with Uncertainty, (ECSQARU'07)*, pp.853–864.

COMPAQH project (2008) *Coordination pour la Mesure de la Performance et l'Amélioration de la Qualité Hospitalière*, available at http://ifr69.vjf.inserm.fr/compaqh/.

de Santis, L., Scannapieco, M. and Catarci, T. (2003) 'Trusting data quality in cooperative information systems', *Proceedings of the International Conference on Cooperative Information Systems (CoopIS'03)*.

English, L. (1999) *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, John Wiley & Sons, Inc.

Etcheverry, L., Peralta, V. and Bouzeghoub, M. (2008) 'Qbox-foundation: a metadata platform for quality measurement', *Proceeding of the 4th Workshop on Data and Knowledge Quality (QDC'08)*.

Fitzpatrick, R. (1996) *Software Quality: Definitions and Strategic Issues*, Staffordshire University, School of Computing Report.

Jankowka, M.A. (2000) *The Need for Environmental Information Quality: Issues in Science and Technology Librarianship*, available http://www.library.ucsb.edu/istl/00-spring/article5.html.

Jarke, M. and Vassiliou, Y. (1997) 'Data warehouse quality: a review of the DWQ project', *Proceedings of the 2nd International Conference on Information Quality (IQ'97)*.

Jarke, M., Jeusfeld, M.A, Quix, C. and Vassialiadis, P. (1999) 'Architecture and quality in data warehouses: an extended repository approach', *Information Systems*, Vol. 24, No. 3, pp.229–253.

Kerr, K. and Norris, T. (2008) 'Improving health care data quality: a practitioner's perspective', *International Journal of Information Quality*, Vol. 2, No. 1, pp.39–59.

Malak, G., Badri, L., Badri, M. and Sahraoui, H. (2004) 'Towards a multidimensional model for web-based applications quality assessment', *Proceedings of the 5th International Conference on E-Commerce and Web Technologies (EC-Web'04), Lecture Notes in Computer Science 3182*, Springer, pp.316–327.

MEDLINE (2010) *U.S. National Library of Medicine's Bibliographic Database and MEDLINE's Proprietary Medical Subject Headings (MeSH) Thesaurus*, available at http://www.proquest.com/en-US/catalogs/databases/detail/medline_ft.shtml.

Miettinen, M. and Korhonen, M. (2008) 'Information quality in healthcare: coherence of data compared between organization's electronic patient records', *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS'08)*, pp.488–493.

Moody, D. (2005) 'Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions', *Data & Knowledge Engineering*, Vol. 55, No. 3, pp.243–276.

Müller, H. and Naumann, F. (2003) 'Data quality in genome databases', *International Conference on Information Quality (IQ 2003)*, pp.269–284.

Naumann, F. and Rolker, C. (2000) 'Assessment methods for information quality criteria', *Proceedings of the International Conference on Information Quality (IQ'2000).*

Naumann, F., Freytag, J.C. and Leser, U. (2003) 'Completeness of information sources', *Proceedings of the International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03).*

Naumann, F., Leser, U. and Freytag, J.C. (1999) 'Quality-driven integration of heterogeneous information systems', *Proceedings of the International Conference on Very Large Data Bases (VLDB'99).*

Oasis (2008) *Web Services Quality Model v1.0*, Committee Draft, 28 November, available at http://www.oasis-open.org/committees/download.php/29803/07.%20wsqmws_quality_model-cd-v1.0-r01.doc.

Peralta, V. (2006) *Data Quality Evaluation in Data Integration Systems*, PhD thesis, Université de Versailles, France and Universidad de la República, Uruguay, available at http://tel.archives-ouvertes.fr/docs/00/32/51/39/PDF/these.pdf.

Pipino, L.L., Lee, Y.W. and Wang, R. (2002) 'Data quality assessment', *Communications of the ACM*, Vol. 45, No. 4.

QUADRIS project (2009) *Quality of Data and Multi-source Information Systems*, available at http://deptinfo.cnam.fr/xwiki/bin/view/QUADRIS.

Rahm, E. and Do, H.H. (2000) 'Data cleaning: problems and current approaches', *IEEE Data Engineering Bulletin*, Vol. 23, No. 4.

Rawashdeh, A. and Matalkah, B. (2006) 'A new software quality model for evaluating COTS components', *Journal of Computer Science*, Vol. 2, No. 4.

Redman, T. (1996) *Data Quality for the Information Age*, Artech House Inc.

Salanti, G., Sanderson, S. and Higgins, J. (2005) 'Obstacles and opportunities in meta-analysis of genetic association studies', *Genetics in Medicine*, Vol. 7, No. 1, pp.13–20.

Strong, D., Lee, Y. and Wang, R. (1997) 'Data quality in context', *Communications of the ACM*, Vol. 40, No. 5.

US Environment Protection Agency (2004) *Increase the Availability of Quality Health and Environmental Information*, available at http://www.epa.gov/oei/increase.htm (accessed on August 2004).

van Deursen, T., Koster, P. and Petkovic, M. (2008) 'Hedaquin: a reputation-based health data quality indicator', *Electronic Notes in Theoretical Computer Science*, Vol. 197, No. 2, pp.159–167.

Wang, R. and Strong, D. (1996) 'Beyond accuracy: what data quality means to data consumers', *Journal on Management of Information Systems*, Vol. 12, No. 4, pp.5–34.

Wang, R.Y., Storey, V.C. and Firth, C.P. (1995) 'A framework for analysis of data quality research', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 4, pp.623–640.

Widom, J. (2005) 'Trio: a system for integrated management of data, accuracy, and lineage', *Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR'05)*, pp.262–276.

Notes

1 'N'habite Pas à l'Adresse Indiquée' is the translation for 'return to sender' in French postal organisms.