2009

A utility-based model to define the optimal data quality level in IT service offerings

Cinzia Cappiello
Politecnico di Milano, cappiell@elet.polimi.it

Marco Comuzzi
City University London, sbbd286@soi.city.ac.uk

A UTILITY-BASED MODEL TO DEFINE THE OPTIMAL DATA QUALITY LEVEL IN *IT* SERVICE OFFERINGS

Cinzia Cappiello, Politecnico di Milano, Dipartimento di Elettronica e Informazione, Via Ponzio 34/5, 20133 Milano, Italy, cappiell@elet.polimi.it

Marco Comuzzi, City University London, Department of Computing, School of Informatics, Northampton Square, EC1V 0HB London, UK, sbbd286@soi.city.ac.uk

Abstract

In the information age, enterprises base or enrich their core business activities with the provision of informative services. For this reason, organizations are becoming increasingly aware of data quality issues, which concern the evaluation of the ability of a data collection to meet users' needs. Data quality is a multidimensional and subjective issue, since it is defined by a variety of criteria, whose definition and evaluation is strictly dependent on the context and users involved. Thus, when considering data quality, the users' perspective should always be considered fundamental. Authors in data quality literature agree that providers should adapt, and consequently improve, their service offerings in order to completely satisfy users' demands. However, we argue that, in service provisioning, providers are subject to restrictions stemming, for instance, from costs and benefits assessments. Therefore, we identify the need for a conciliation of providers' and users' quality targets in defining the optimal data quality level of an informative service. The definition of such equilibrium is a complex issue since each type of user accessing the service may define different utilities regarding the provided information. Considering this scenario, the paper presents a utility-based model of the providers' and customers' interests developed on the basis of multi-class offerings. The model is exploited to analyze the optimal service offerings that allow the efficient allocation of quality improvements activities for the provider.

Keywords: Data Quality, Efficiency, Service Offerings.

1 INTRODUCTION

Enterprises increasingly focus their business strategy on information management, since an effective use of organizational data can have a considerable impact on business decisions and provide high benefits. Furthermore, information can be often exploited to offer additional and valuable services to external customers. Informative services manipulate raw data and produce *information products* (Ballou et al. 1998). In order to evaluate the service effectiveness and thus the ability of the provided information to meet both the organizational and users' (customers) requirements, it is then possible to consider data quality theories. Data quality is a multi-dimensional concept that evaluates the suitability of data to the tasks for which they are required, and thus to the users that access them. Data quality can be assessed by means of different dimensions, whose definition and evaluation are strictly dependent on the context and users involved. For this reason, the users' perspective has been always considered fundamental in data quality. Consequently, literature contributions have always focused their attention on the definition of methodologies and methods that support providers in the achievements of data quality targets that would completely meet users' needs. Quality management mainly suggests the adoption of the *Zero Defects* approach that consists in setting targets to the highest quality values (English 1999). However, if the organization follows a zero defects approach in areas which do not need it, resources may be wasted. Moreover, reaching the highest quality values might lead to quality improvement that the organization may not be able to afford. In fact, it is necessary to consider that

providers have their own requirements in provisioning services and many times the complete satisfaction of users' requirements is not convenient since costs are greater than benefits. Hence, the Zero-Defects approach to data quality management is often excessive, since it does not consider that the data quality improvement is not a trivial task and in some cases it requires very expensive projects, which are not always feasible for the service providers. Generally, we argue that it would be better to adopt an approach that fixes data quality targets on the basis of a conciliation of providers' and users' needs. The definition of this equilibrium is a complex issue since, for each provided information, we can have different utilities depending on the type of user that accesses it. Considering this scenario, the paper presents a utility-based model of the providers' and customers' interests developed on the basis of multi-class offerings. The model is exploited to analyze the optimal service offerings that allow the efficient allocation of quality improvements activities for the provider.
The paper is organized as follows. Section 2 reviews the literature on similar contributions. Section 3 presents the main useful concepts for data quality management and shows the model for the definition of the users and providers quality targets. Section 4 presents the model of the provider and users' utility functions in a data service scenario and discusses the issue of optimal data quality level definition in informative service offerings.

2 RELATED WORK

The identification of service offerings that define the most suitable quality targets that contemporarily satisfy providers and users' needs is a research issue that can be generally related to the identification of quality level agreements. This is a new open issue in the data quality field, as well as in the broader field of Service Oriented Computing. Here, the Service Level Agreement (SLA) is defined as a binding contract which formally specifies end-user expectations about the solution and tolerances, i.e., it is a collection of service level requirements that have been negotiated and mutually agreed upon by the information providers and the information consumers. In fact, providers define some service levels as a fixed combination of their specific capabilities on a set of quality dimensions that are also considered by the users to define their targets. If providers' capabilities and users' needs are not immediately compatible, a negotiation phase is required in order to find the most suitable conciliation between providers' and users' quality targets. In this field, there are several languages proposed for the definition and monitoring of the SLA such as WSLA (Keller and Ludwig 2002) or WS-Agreement (Ws-Agreement Framework 2003). WSLA allows providers to define quality dimensions and to describe functions to evaluate them. Furthermore, it provides monitoring of the parameters during operations and invocation of recovery actions when contract violations occur. Similarly, WS-Agreement provides constructs for advertising the capabilities of providers and creating agreements based on creational offers, and for monitoring agreement compliance at runtime. Once the service capabilities description is provided, the selection of the most suitable service is enabled by the definition of the users requirements. The SLA definition starts from provider capabilities and users' requirements specification and defines all the condition of the service provisioning.
In the data quality field, quality requirements are focused on a set of criteria able to define the suitability of a data set for the process in which it is involved. Data quality is a multi-dimensional concept and the data quality literature provides a thorough classification of data quality dimensions, even if there are discrepancies on the definition of most dimensions due to the contextual nature of quality. The six most important classifications are presented in (Wand and Wang 1996, Wang and Strong 1996, Redman 1996, Jarke et al. 1999, Bovee et al. 2001, Naumann 2002). By analyzing these classifications, it is possible to define a basic set of data quality dimensions including accuracy, completeness, consistency, timeliness, interpretability and, accessibility, which represent the dimensions considered by the majority of the authors (Scannapieco and Catarci 2002). The assessment of these dimensions reveals the ability of a data collection to meet users' needs.
In the literature, data quality users' requirements have been mostly used as one of the driver for the identification of the most suitable data source (e.g., Scannapieco et al. 2004). Users' requirements have been sometimes translated into utility functions. In (Even and Shankaranarayanan 2007), utility

functions have been used by supporting multiple assessments of quality, each within a different usage context. Utility functions have been also used to alleviate the problem of data fusion in the presence of inconsistencies, for example in combining different versions of the same data (Motro et al. 2004).
As already discussed in the Introduction, in the data quality field, the provider perspective has been scarcely considered. Data quality agreements issue has been only addressed in quality-constrained data provisioning (Missier and Embury 2005). Missier and Embury (2005) propose a framework for the definition of formal agreements between the provider and the customers. Focusing on the completeness dimension, they also provide an algorithm for dealing with constraints on the completeness of a query result with respect to a reference data source.
In our work, the approach can be considered innovative since providers capabilities are not fixed at priori. In fact, we primarily consider the users' requirements and we assume that the provider capabilities are functions of the current quality level of their IT services and of the costs related to the improvement activities needed to satisfy users requirements.

3 THE SERVICE PROVISION AND DATA QUALITY REQUIREMENTS SPECIFICATION

A business process can be composed and executed by means of IT and physical services. The former are services that are responsible for data manipulation and that aim at generating and providing useful information. The latter are business services that are composed of physical activity that cannot be made automated (e.g. delivery of goods). In this paper we focus on the first type of services and we characterise them by considering functional and non-functional requirements. Since the output provided by IT services is *information*, the quality of such services can be mainly evaluated by considering data quality dimensions.
In our model, in the data quality assessment phase, we consider the quality dimensions introduced in the previous section and define an aggregate measure of data quality level (*qc*) by using a weighted average, that is:

$$qc = \sum_{i=1}^{N} w_i \cdot dq_i \qquad (1)$$

where w_i are the weights that denote the importance of the single dimension dq_i for the user or the provider and N is the total number of the considered criteria. In order to use this model, we make the main assumption to consider the quality dimensions independent of each other.
If the assessment results reveal that the provider sources are characterized by an insufficient data quality level, the adoption of quality improvement techniques should be considered. Improvement methods are distinguishable in *data-oriented* and *process-oriented* techniques. The former focus on error detection and correction, whereas the latter aim at identifying and correcting the activity in the process responsible for the error. Therefore, data-oriented techniques are characterized by low investment costs and short term benefits, whereas process-oriented techniques imply a very high investment cost, even though they are likely to provide long-term benefits. Process-oriented techniques are, in general, to prefer, since data-oriented techniques need to be performed periodically to obtain long-term benefits and thus the total cost will be higher than the initial investment of any process-oriented technique.
In the framework proposed in this paper, the providers should evaluate their convenience to improve the data quality level by also considering that low data quality levels raise poor quality costs, mainly due to service failures and consequent repair actions.
A fundamental hurdle is that costs and benefits of attaining a certain data quality level are difficult to estimate ex ante. We consider a distinction between *non-quality* and *quality* costs.

- *Non-quality costs:* they are costs associated with poor data quality and, consequently, with all the activities necessary to correct errors and re-execute tasks.
- *Quality costs:* they are associated to the activities and resources necessary in the improvement project.

Improvement interventions may be of variable complexity. They could regard the purchase and implementation of standard software tools (e.g., data cleaning tools), the design and the development of ad-hoc software modules or, in the most complex cases, the re-organization of the whole IT architecture (e.g. for improvement of information availability or security). Therefore, quality costs can include licence costs, hardware costs for the acquisition of new machines and human resources costs for analysis, development or implementation activities.
It is necessary, however, to consider that non-quality costs can be considered as a potential saving, and represent tangible benefits of quality improvement. The benefits of the improvement process are at least equal to the savings from non-quality costs. Additional tangible and intangible benefits can be achieved in higher-performance scenarios. It must be noted that the quality costs depend on the improvement techniques that are implemented and benefits are related to the type of services that are improved through data enhancement.

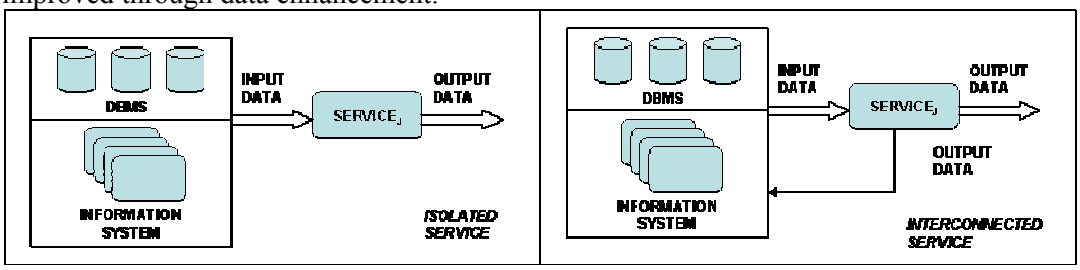

Figure 1 - Isolated and interconnected services

On the basis of the role played by information, it is possible to distinguish different types of IT services. In this paper, we consider a distinction between isolated and interconnected services (see Figure 1):

- *Isolated services*: these are services for which the output data are specially produced for the final user and they are not used in the information system of the organization, which supports the execution of the organization's business processes.
- *Interconnected Services*: they are services which produce information that is also used in the organization's daily operational activities.

Systems that provide isolated services can be compared to open loop systems in which the improvements in the quality of output data are totally dedicated to meet the user requirements. In fact, such improvements do not impact on the provider operational processes and benefits from data quality improvements will derive only by the increase of the customer satisfaction. Conversely, systems that provide interconnected services can be compared to closed loop systems in which improvements are likely to influence the organization's internal business processes and will produce higher benefits for the provider. In fact, improvements of the output data will also impact on the correctness of operational data and thus, on the execution of all the business processes. In this case, improvements decrease the probability that services might fail as well as the poor quality costs.
Real-time data about stock quote rates, for instance, can be provided by either financial brokering institutions or merchant bankers. In the former case, we can label the service as isolated, since brokers simply collect data from different sources in order to satisfy the requirements of their customers. In the latter case, the stock quote provisioning service can be considered interconnected, since financial institutions, besides selling data to customers, are also likely to exploit the same data for their internal activities, e.g., managing customers' investment portfolios.
Generally, we can introduce a coefficient α to express the degree of interconnection of a service. It defines the impact that a quality improvement intervention undertaken for a service has on the provider operational processes. If $\alpha = 0$, then the service provisioning process is not connected with the organizations business processes, whereas if $\alpha > 0$ the service provisioning process shows an increased degree of interconnection with internal business processes ($0 \leq \alpha \leq 1$) .
Furthermore, it is also necessary to clarify that service providers are used to offer services along different quality configurations (i.e., service levels) in order to satisfy different user requirements. As

an example, we can consider the visualization of stock values in a trading activity, a classical financial service for which timeliness is a critical dimension. For the sake of simplicity, let us consider two classes of users: *ordinary retail customers* and *traders*. The two classes of users have different requirements on stock information. The ordinary retail users can tolerate a lower quality than the trader, as they usually access the service with a lower frequency. Moreover, errors on stock values have a considerably lower impact, since operations performed by ordinary retail users are usually less risky and involve lower amount of money. On the other hand, the trader needs accurate and updated information and, therefore, requires high-quality information. The effects of an inaccurate or delayed value can be disastrous, since the trader is often involved in very risky operations involving stocks from a large portfolio of customers. Therefore, banks are likely to provide two different quality profiles for the stock quotes information service. The first profile is characterized by an acceptable value of timeliness while the other one specifies real-time information provisioning and is also associated with a higher cost.

In respect of this example, we consider an information service associated to several classes of users. Users belonging to a class have the same requirements on the quality of the data provided by the service S. In details, our model of service offerings considers a service S and assumes that a user (or customer) u is assigned to one of the K user classes UC_k, where $k=1,\ldots,K$. Each class contains users with similar characteristics. The number of users in a given class UC_k is indicated as M_k. Users belonging to the same class are associated with the same quality requirements for the service S. For each user class UC_k, we define the data quality level qc_k defined in the service offerings $QC(S)$ for service S. Note that each qc_k is calculated as a weighted average of the requirements specified for the different quality dimensions by using the formula shown in Eq. 1. Hence, the service offering QC(S) for service S is defined as a set of increasing data quality levels associated to K classes, that is:

$$QC(S) = \{qc_1, \ldots, qc_K\}. \quad (2)$$

From the provider perspective, the aim is to define a service offering $QC(S)$ that satisfies some optimization criteria. A first criterion can be of defining service offerings on the basis of the fulfilment of the user requirements. Usually, such criteria tend to minimize the specification of subjective quality levels, since service offerings are developed to best fit user requirements. In the next section we introduce a utility model for describing the provider and the customers' interest, and define a criterion for defining service offerings which jointly considers the interests of both the provider and the customers.

4 A MODEL FOR SERVICE OFFERINGS EFFICIENCY ASSESSMENT

In order to define an efficient way for the provider to define service offerings and to decide the quality improvement actions to be performed on data, we first need to introduce a model which defines the provider and the customers' utility functions in our informative service scenario.

The model relies on the definition of utility functions for both the data provider and customers. In our model, we adopt quasi-linear utility functions (Jackson 2003). Quasi-linear utility functions represent an efficient and compact modeling tool for negotiations and bargaining problems in which it is easy to isolate, for every participant, *value* and *payment* terms. We argue that the case of data quality and, specifically, information service offerings falls within such category. Sources of benefits and costs related to data service offerings for providers and customers, in fact, have already been analyzed by a large body of academic literature (Batini et al. 2006, Eppler and Helfert 2004, English 1999, Loshin 2001].

Quasi-linear utility functions are such that the utility value for an agent on a given contract X is defined by two terms, i.e., a value and a payment term. The value term determines the value obtained by an agent from the contract, whereas the payment term refers to the amount of money that an agent is going to receive or pay for the contract. Value and payment terms can be either positive or negative.

For the provider, the payment term is positive and value term is negative, because the provider receives money from customers, but, at the same time, it sustains a cost for providing the negotiated contract, therefore losing utility. Conversely, the payment term is negative for customers, whereas the value term is positive, because the customers pay money for a contract and, at the same time, have a positive evaluation of the contract negotiated with the provider.
We first introduce the definition of quasi-linear utility functions for data providers and customers in the multi-class data service scenario introduced in the previous section. Then, we show how the utility model can be exploited to provide a criterion for the provider to define the optimal service offerings and, consequently, clarify which quality improvements need to be performed. The optimal service offering obtained through our model is then compared against the findings of the zero-defect approach for data quality management, which, generally, implies the complete fulfillment of the customers' quality requirements.
Generally, a quasi-linear utility function defined for an agent *P* behaving in a service provider's perspective is defined as:

$$U_P = Price(X) - Cost(X) \tag{3}$$

where *Price*(X) is the price, that is, the amount of money obtained by *P* for providing the generic contract *X* (payment term), whereas *Cost*(X) represents the cost sustained by *P* to provide the contract X (negative value term).
Similarly, for the generic agent *C* behaving as a service customer, the utility of a contract U_C(X) assumes the following form:

$$U_C = Value(X) - Price(X) \tag{4}$$

where *Value*(X) is the value generated by the contract X to the customer, whereas *Price*(X) is, as in U_P(X), the price that the customer has to pay for receiving the contract X.
According to the model presented in Section 3, in our multi-class data service scenario, the contract X assumes the form of the service offering *QC*:

$$X = QC = (qc_1, \ldots, qc_K) \tag{5}$$

where qc_k is the data quality level provided to customers in class *k*, with $k = 1,\ldots K$.
In the following, we will consider two agents, i.e. the provider *P*, providing the service *S*, and the collection of customers *C* of service *S*.
The total amount of money *Price*(*QC*) received by the data provider *P* for the provisioning of a given service offering *QC* is given by the sum of the money received from customers in each service class defined in the service offering, that is:

$$P(QC) = \sum_{k=1}^{K} p(qc_k) \cdot M_k \ , \tag{6}$$

where $p(qc_k)$ is the price of data provided for users in class *k*, while M_k is the number of users that belong to class *k*.
The term *Cost*(*QC*) represents the cost sustained by the provider to provide a service offerings *QC*. and can be expressed as:

$$Cost(QC) = C_P(qc_K) - B_P(qc_K) \tag{7}$$

where $C_P(qc_K)$ is the cost actually sustained by the provider to provide the maximum quality level qc_K to its customers, whereas $B_P(qc_K)$ is a quantification of the benefits introduced in the provider's internal business processes by attaining a maximum data quality level qc_K. $C_P(qc_K)$, is a function of only the maximum data quality level qc_K in *QC*(*S*) since we argue that, once the provider commits to

provide qc_K to its customers, the marginal cost of providing the data service with a lower quality level qc_k, with $1 < k < K$-1, is negligible. Similarly, we also argue that it is rational for the provider to use, in its internal business processes, data at the maximum quality level qc_K. Therefore, $B_P(qc_K)$ is a function of only the maximum data quality level qc_K.
The value term $Value(QC)$ in $U_C(QC)$ can be expressed as the sum of values $v_k(qc_k)$ generated for each class of customers k. Therefore:

$$Value(X) = Value(QC) = \sum_{k=1}^{K} v_k(qc_k) \tag{8}$$

Our objective is to study the optimal service offering $QC^*=(qc_1^*,\ldots,qc_K^*)$ that maximises the sum of the utility for the provider and the service customers. Moreover, we demonstrate that such QC^* differs from the optimal QC^z imposed by the zero-defect approach to data quality management, which implies the service offering to fully satisfy the requirements of the customers. More specifically, in our model, we argue that the full satisfaction of customers' requirements occurs when a service offering maximises the value $v_k(qc_k)$ for each class of customers. Therefore, in the zero-defect approach, the optimal service offering QC^z can be defined as follows:

$QC^z = \left(qc_1^z,\ldots,qc_K^z\right)$, where $qc_i^z = \arg\max v_i(qc_i)$ for $i=1,\ldots,K$.

The sum of the utility of customers and suppliers can be expressed:

$$U_P(QC) + U_C(QC) = Value(QC) - Cost(QC) = \sum_{k=1}^{K} v_k(qc_k) - C_P(qc_K) + B_P(qc_K). \tag{9}$$

Since U_P+U_C, as a function of QC, is separable in the variables qc_k, with $k=1,\ldots,K$, the conditions under which $U_P(QC)+U_C(QC)$ is maximised can be expressed as:

$$\left.\frac{\partial(U_P + U_C)}{\partial qc_k}\right|_{k=1\ldots K} = 0 \tag{10}$$

Since both C_P and B_P are a function of only qc_K, for the first K-1 equations implied by Eq. 10, we can write the following:

$$\left.\frac{\partial(U_P + U_C)}{\partial qc_k}\right|_{k=1\ldots K-1} = \frac{dv_k(qc_k)}{dqc_k} = 0 \tag{11}$$

While the Kth equation becomes:

$$\left.\frac{\partial(U_P + U_C)}{\partial qc_k}\right|_{k=K} = \frac{dv_K(qc_K)}{dqc_K} - \frac{dC_P(qc_K)}{dqc_K} + \frac{dB_P(qc_K)}{dqc_K} = 0\,. \tag{12}$$

In order to go into detail in the analysis of the optimal service offering QC*, we need to characterise the functions $v_k(qc_k)$, $C_P(qc_K)$, and $B_P(qc_K)$ in our information service offering settings.
Considering past contributions (Batini et al. 2008, Eppler and Helfert 2004, English 1999, Loshin 2001) on data quality costs, it is possible to consider the following form of costs $C_P(qc_K)$ sustained by a provider:

$$C_P(qc_K) = F + P \cdot qc_K + I \cdot \exp(qc_K) \tag{13}$$

where the F, P, and I are coefficients that reflect the complexity of a generic data quality project. The coefficient F is related to the fixed costs of the data quality projects required for achieving qc_K, such as licence costs or hardware costs for the software and IT infrastructure required by the project. The coefficient P relates to the project development part, in which we have variable costs associated to the analysis and implementation activities that are evaluated by considering their duration and involved human resources. The coefficient I is related to the improvement of some data quality dimensions (e.g., availability, accessibility, security). Such improvements usually require considerable changes in the whole IT architecture and therefore the costs associated to them grows exponentially with the level of data quality qc_K that needs to be achieved.

We make the assumption that the benefits obtained by the service provider to attain a maximum data quality level qc_K are a fraction α, with $0 < \alpha < 1$ of the sustained costs. Such a fraction α is determined by the degree of interconnection of the provider's internal business processes defined in Section 3. Specifically, the higher the degree of interconnection, the higher the benefits that the provider can obtain on its internal business processes. Therefore:

$$B_P(qc_K) = \alpha \cdot C_P(qc_K) = \alpha F + \alpha P \cdot qc_K + \alpha I \cdot \exp(qc_K) \tag{14}$$

We propose two different types of value functions $v_k = v_k(qc_k)$ for the set of customers in the generic k-th class.

Type 1: Gaussian customers value functions.

In this case (see Figure 2), the optimal data quality value qc_k^z for customers in the kth class under the zero-defect principle is $qc_k^z = \mu_k$, that is, the one that maximises the customers' value.

$$v_k(qc_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(qc_k - \mu_k)^2}{\sigma_k^2}}$$

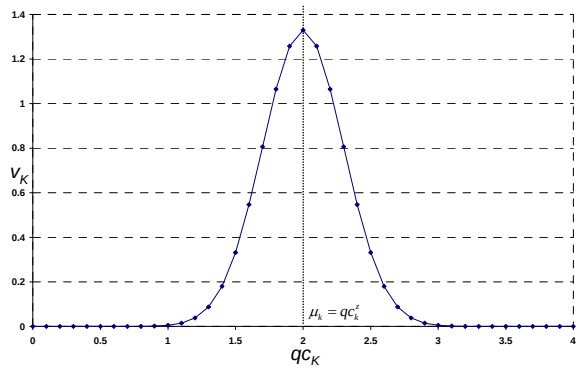

Figure 2 - Gaussian utility function for customers in class UC_k

Type 2: Monotonic increasing customers value functions.

The second type of value functions considers monotonic increasing value that saturates at a certain value qc^z (specifically, we use a sigmoid function to express this second type of value function, see Figure 3). Such level qc is the one identified by the zero-defect approach as the optimal data quality level for customers in the k-th class, since it maximises the customers' value.

$$v_k(qc_k) = \frac{K}{1 + e^{-qc_K}} - 0.5$$

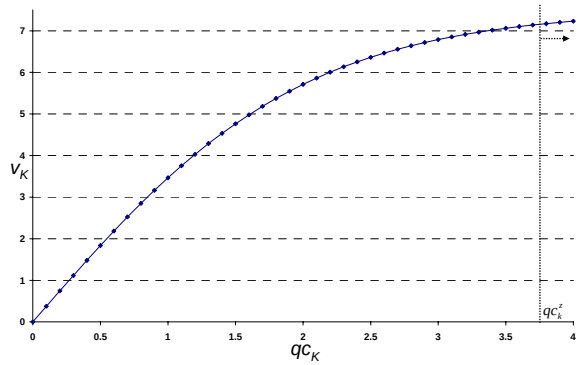

Figure 3 - Increasing Monotonic utility function for customers in class UC_k

The Gaussian increasing value function is suitable to model cases in which customers cannot accept higher or lower quality values than the requested ones. Considering the example discussed in Section 3, normal customers can accept out-of-date stock values since they do not have to use these data in critical processes, but they cannot accept higher quality level, since this would require an additional cost that they should pay for information that they do not actually need. The monotonic increasing value function is suitable to model cases in which customer define their acceptable quality level as the minimum acceptable value. For example, traders obviously cannot accept out-of-date stock values and, therefore, they are likely to fix a minimum quality requirement. However, traders are also likely to be equally satisfied with a quality level that exceeds their minimum requirements, since they may need to deal with unpredictable and critical situations which could be benefit from higher quality of data.

In respect of Eq. 11, the optimal data quality level in our model qc_k*, with k=1,…,K-1, coincides with the optimal data quality model identified by the zero defect approach, that is, $qc_k^* = qc_k^z$.

In fact, for both Gaussian and monotonic increasing value functions, the following condition holds:

$$qc_k^* = \left\{ qc_k : \frac{dv_k(qc_k)}{dqc_k} = 0 \right\} = qc_k^z \text{ for k=1,…,K-1.}$$

In other words, both our model and the zero-defect approach to data quality management imply that customers who do not ask for the maximum level of data quality should be provided with a level of data quality that fully satisfy their requirements, that is, that maximises their value.

For the maximum level of quality provided to customers, the findings of our model differ from the corresponding findings obtained with the zero-defect approach, i.e. $qc_K^* \neq qc_K^z$.

In our model, the data quality level qc_K is determined by solving Eq. 12, which can be rewritten as:

$$\frac{\partial v(qc_K)}{\partial qc_K} - P(1-\alpha) - I\exp(qc_K) + \alpha \cdot \exp(qc_K) = 0\,; \tag{15}$$

which then leads to the following equation:

$$\frac{\partial v(qc_K)}{\partial qc_K} = P(1-\alpha) + I \cdot (1-\alpha) \cdot \exp(qc_K)\,. \tag{16}$$

A first consideration that must be made is that the maximum level of data quality qc_K does not depend on the fixed costs F of the data quality project. This is consistent with the fact that, by definition, fixed costs must be sustained for any data quality project, the maximum level of attained quality qc_K notwithstanding.

Eq. 16 can be solved graphically in two cases C1 and C2 that consider, respectively, Gaussian and monotonic increasing functions for modelling the customers' value. A graphical representation of the solutions of Eq. 16 in case C1 and C2 is given in Figure 4 and Figure 5, respectively. Please note that in the graphical solution, the function $f(x)$ represents the derivative of the customers' utility function on $v(qc_K)$ (i.e. left argument in Eq. 16).

The graphical representation of the solution to Eq. 16 in case C1 is reported in Figure 4.

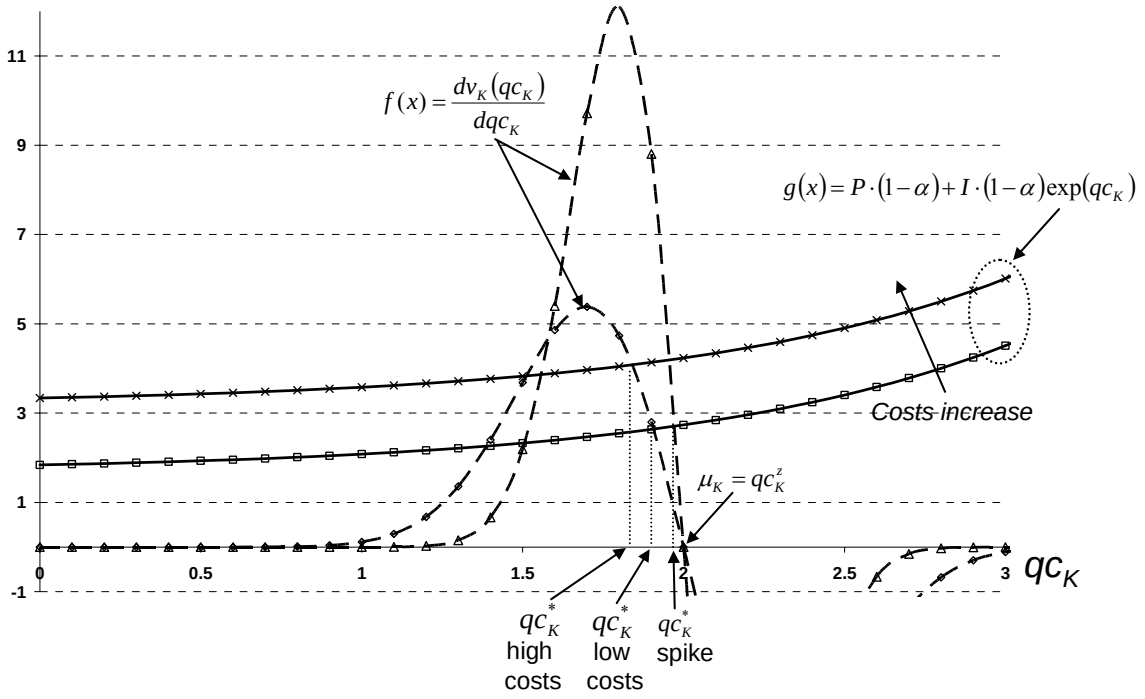

Figure 4 - Determining the optimal data quality level (Case C1, Gaussian customer utility)

The first conclusion that can be drawn is that, generally, $qc_K^* < qc_K^z$, that is, our model implies a maximum data quality level that is lower than the optimal level predicted adopting the zero-defect approach. This is consistent with the fact that, while the zero-defect approach only considers the maximisation of customer's requirements, our model considers a trade-off between the costs sustained by the provider to provide a certain maximum data quality level (and the benefits induced by these) and the value produced for the customers requiring maximum data quality.

Second, the maximum data quality level qc_K^* decreases as the costs (both P and I) for providing a certain data quality level increase (see qc_K^* *low costs* and qc_K^* *high costs* in Figure 4). Moreover, qc_K^* decreases as α decreases, since both terms $P(1-\alpha)$ and $I(1-\alpha)$ increase. A lower value of α implies a lower degree of interconnection, as defined in Section 3, of the provider's business processes. In our modelling of the informative service provisioning scenario, a lower degree of interconnection implies a decrease of the benefits induced on internal business processes by the attainment of a certain level of data quality. Therefore, our model can be used to identify which is the impact of the degree of interconnection on the decrease of the maximum data quality level qc_K^* that can be provided to customers.

Third, it has to be noticed that the more spike shaped the Gaussian function representing the value created for customers, the closer the quality level qc_K^* to the zero-defect approach level qc_K^z (see qc_K^* *spike* and qc_K^* *low costs* in Figure 4). This can be interpreted by stressing that when customers' interested are clearly stated, i.e., the function representing the value created for customers in the Kth class is spiked around μ_K, then it will be easier for the provider to identify the optimal trade-off between the costs sustained for providing quality of data and the value generated for customers (and this will be closer to the optimal satisfaction of customers' requirements hypothesized by the zero defect approach to data quality management).

Finally, in case C1, it has to be noticed that when costs are too high, Eq. 16 cannot be solved, i.e., there is no intersection between the two functions (dotted and straight lines in Figure 4, respectively)

represented in the Figure 4. In other words, it is impossible for the provider to find the trade-off between the cost of quality of data and the value generated for customers.

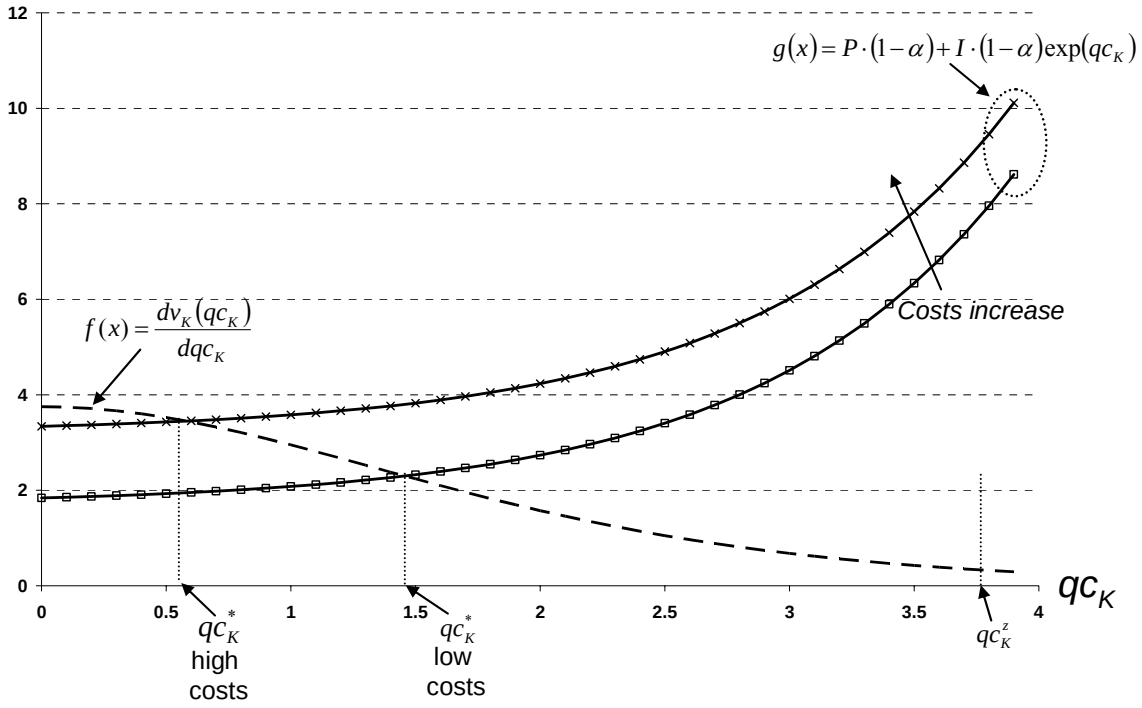

Figure 5 - Determining the optimal data quality level (Case C2, Monotonic increasing customer utility)

The graphical representation of Eq. 16 in case C2 is shown in Figure 5. The major findings already discussed for case C1 hold also in this case. More in detail, the maximum data quality level qc_K^* predicted by our model will always be lower than the value qc_K^z implied by the zero-defect approach and qc_K^* decreases as (*i*) the costs sustained for providing data with a certain level of quality increases and (*ii*) the provider's business process are less interconnected (i.e., α decreases).
Similarly to what happens in C1, a solution to Eq. 16 may not be found if costs are very high also in C2. Moreover, it has also to be noticed that the value qc_K^* drastically decreases as the cost terms $P(1-a)$ and $I(1-a)$ increase.

5 CONCLUDING REMARKS AND FUTURE WORK

The paper has presented a novel approach for defining optimal service offerings for information services. In particular, our model defines the optimal service offering as the one that maximises the sum of the service provider and customers' utility functions. The optimal service offering obtained with our model differs by the one defined by the zero-defect approach in the definition of the maximum quality level. In particular, our model argues for lower maximum quality levels, in order to keep into consideration the trade-off between the costs sustained by the provider for improving the quality of data, the value created by the service offering for customers, and the benefits obtained by the provider on its internal business processes from the improvement in the quality of data.

The limitations of our model imply the need for future work on the model development. First, the model relies on the ability of the service provider to estimate the utility functions of the classes of users for the provided service. Understanding preferences of users requires the development of user profiling and clustering techniques, which should be further investigated. Second, our model defines

optimality of service offerings in terms of the maximisation of the sum of the provider and customers utilities. Optimality may be defined according to other metrics involving utilities of the involved actors. In particular, we want to investigate the notion of equilibrium of service offerings, i.e., studying optimal service offerings that define an equilibrium among the service provider and the customer in the utility space.

References

Ballou D. P., Wang R., Pazer H.L. and Tayi G.K. (1998) Modelling Information Manufacturing Systems to Determine Information Product Quality. Management Science, 44 (4).

Batini C., Cabitza F., Cappiello C., and Francalanci C. (2008) A Comprehensive Data Quality Methodology for Web and Structured Data. International Journal of Innovative Computing and Applications, 1(3), 205-218.

Bovee M., Srivastava R.P. and Mak, B. (2001) A Conceptual Framework and Belief- Function Approach to Assessing Overall Information Quality. Proceedings of the ICIQ ‘01.

English L. (1999) Improving Data Warehouse and Business Information Quality. John Wiley & Sons.

Eppler M.J. and Helfert M. (2004) A Framework For The Classification Of Data Quality Costs And An Analysis Of Their Progression. Proceedings of ICIQ '04, 311-325.

Even, A. Shankaranarayanan G. (2007) Utility-driven assessment of data quality. Data Base, 38(2), 75-93.

Jackson M.O. (2003) Mechanism Theory. In the Encyclopaedia of Life Support Systems, edited by Ulrich Derigs, EOLSS Publishers.

Jarke, M., Jeusfeld, M.A., Quix, C., Vassiliadis, P. (1999) Architecture and Quality in Data Warehouses: an Extended repository Approach. Information Systems, 24 (3).

Keller, A., Ludwig, H. (2002) The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services. Technical Report RC22456(W0205-171), IBM Research Division, T.J. Watson Research Center.

Loshin D. (2001) Enterprise Knowledge Management: The Data Quality Approach, Morgan Kaufmann Publishers.

Missier P., Embury S. M. (2005) Provider issues in quality-constrained data provisioning. Proceedings of the International Workshop on Information Quality in Information Systems (IQIS'05).

Motro A., Anokhin P., and Acar A.C. (2004) Utility-based resolution of data inconsistencies. Proceedings of the IQIS '04.

Naumann, F. (2002) Quality-Driven Query Answering for Integrated Information Systems. LNCS 2261.

Orr, K. (1998) Data Quality and Systems Theory. Communications of the ACM, 41(2).

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

Scannapieco, M., Catarci, T. (2002) Data Quality under a Computer Science Perspective. Archivi & Computer (in Italian).

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., Baldoni, R. (2004) The architecture: a platform for exchanging and improving data quality in cooperative information systems. Information Systems, 29(7), 551-582.

Wand, Y., Wang, R. Y. (1996) Anchoring data quality dimensions in ontological foundations. Communication of the ACM, 39(11).

Wang, R.Y. (1998) A Product Perspective on Total Data Quality Management. Communications of the ACM, 41(2).

Wang, R.Y., Strong, D.M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4).

WS-Agreement Framework (2003) https://forge.gridforum.org/projects/graap-wg.