EFFICIENT ALLOCATION OF QUALITY IMPROVEMENT EFFORTS TO SUPPORT THE DEFINITION OF DATA SERVICE OFFERINGS

(Research In Progress)

Cinzia Cappiello
Marco Comuzzi
Politecnico di Milano, Milano, Italy
{cappiell, comuzzi}@elet.polimi.it

The quality of data is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements. The assessment of data quality dimensions should consider the degree to which data satisfy users' needs. User expectations are clearly related to the selected information and at the same time the information can have different utilities depending on the type of user that accesses it. In this paper, the information is considered as a product of a specific service and data quality as a component of the service quality. For each service, it is possible to identify a provider and a final user. In the data quality literature, authors have always only considered as important the final users' perspective declaring that providers should adapt their service offerings in order to completely satisfy users requirements. However, it is necessary to consider that providers have their own requirements in provisioning services since they should evaluate costs and benefits related to their activity. Therefore, we advocate the need for service offerings that define the most suitable quality targets that contemporarily satisfy providers and users' needs. This paper presents a utility-based model of the provider 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.

1. INTRODUCTION

The quality of data is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements [17][21]. Data quality is a multi-dimensional concept that evaluates the suitability of data for the tasks in which they are involved, and thus for the users that access them.

Data quality literature has always focused its 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 Defect approach that consists in setting targets to the highest quality values [7]. However, if the organization follows a zero defects approach in areas which do not need it, resources may be wasted. Furthermore, reaching the highest quality values might lead to quality improvement that the organization may not be able to afford. Hence, this approach is sometimes excessive, since it often involves high quality improvement costs for the service providers. It would be better to adopt an approach that fixes data quality targets on the basis of the requirements of users that access data and of the providers' advantage [5].

In the literature, the providers' perspective has been scarcely considered. In fact, providers have their own requirements in provisioning services since they should evaluate costs and benefits related to their activity. Considering that data quality improvement can raise significant costs, especially if targets to high quality values are fixed, providers should consider the benefits that such improvement activities would

produce. The framework proposed in this paper aims at considering both the users and the providers' perspectives and at providing a support in the definition of optimal service offerings for data services that maximize their gains. Note that the evaluation of the gain implies a utility-based analysis. The goal of the work is the demonstration that the adoption of the Zero Defect approach is often too costly for the providers and useless for the users that have lower data quality requirements.
The paper is organized as follows. Section 2 reviews the data quality literature. Section 3 presents the main useful concepts for data quality management and for the evaluation of costs and benefits and Section 4 shows the model for the definition of the users and providers requirements. Section 5 presents the model that defines the provider and users' utility functions in a data service scenario.

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 but it is a topic investigated in the Service Oriented Computing area. Here, the Service Level Agreement (SLA) is defined as a binding contract which formally specifies end-user expectation about the solution and tolerances. It is a collection of service level requirements that have been negotiated and mutually agreed upon by the information providers and the information consumers. Usually, providers define some service levels as a fixed combination of their specific capabilities on a set of quality dimensions. In this field, there are several languages proposed for the definition and monitoring of the SLA such as WSLA [11] or WS-Agreement [23]. 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 that 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. A framework that supports the automatic generation of the SLA in Web service environment has been proposed in [6].
In the data quality field, the data quality agreements issue has been addressed in quality-constrained data provisioning field [13]. In [13] authors proposes a framework for the definition of a formal agreements between the provider and the customers. Focusing on the completeness dimensions, they also provide an algorithm for dealing with constraints on the completeness of a query result with respect to a reference data source. Utility functions have been instead used to alleviate the problem of data fusion in the presence of inconsistencies, for example in combining different versions of the same data [14].
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 information services and of the costs related to the improvement activities needed to satisfy users requirements. Furthermore, to our knowledge, in the data quality literature there are not previous contributions that address the definition of the optimum data quality level to provide by considering both the customers and provider perspective.

3. THE DATA QUALITY ASSESSMENT AND IMPROVEMENT

The notion of data quality has been widely investigated in the literature. It refers to the degree to which data satisfy user requirements or are suitable for a specific process. Both theoretical and experimental results indicate that data quality is a multi-dimensional concept [1][18][20][22]. 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 & Wang 1996 [20], Wang & Strong 1996 [22], Redman 1996 [18], Jarke 1999 [10], Bovee 2001[4], Naumann 2002 [15]. 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 [19]. Timeliness is usually considered together with other time-related dimensions, typically currency and volatility [1].
In our model, in data quality assessment phase we consider this set of quality dimensions and define an aggregate measure of data quality level (*DQ*) by using a weighted average such as:

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

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 provider sources are characterized by an insufficient data quality level, they should be improved by applying a quality improvement technique. Improvement methods are distinguishable in data-oriented and process-oriented techniques. The former focus in the error detection and correction, while the latter aim at correcting the process that generates the error. Therefore, the former are characterized by low investment costs and short term benefits, while the latter implies 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 are difficult to estimate ex ante. We refer to *non-quality costs* as the costs associated with poor data quality and, consequently, with all the activities necessary to correct errors and re-execute tasks. Instead, *quality costs* are associated with the activities and resources necessary in the improvement project. 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.

4. THE SERVICE PROVISION AND QUALITY REQUIREMENTS SPECIFICATION

On the basis of the role played by information, it is possible to distinguish different types of data services. First of all, it is possible to make a distinction between *informational* and *operational* services. The former are the services for which the information is the output, whereas the latter are services in which the information is used to monitor or control the information flow. This distinction does not influence the data quality requirements, since data should be in any case correct. Anyhow, it is important to notice that there are services for which the information used as output is not used in the information system of the organization. Conversely, other services produce information which is also used in the organization's daily operational activities. This difference can be highlighted as a distinction between *isolated services* and *interconnected services* (see Figure 1**).**

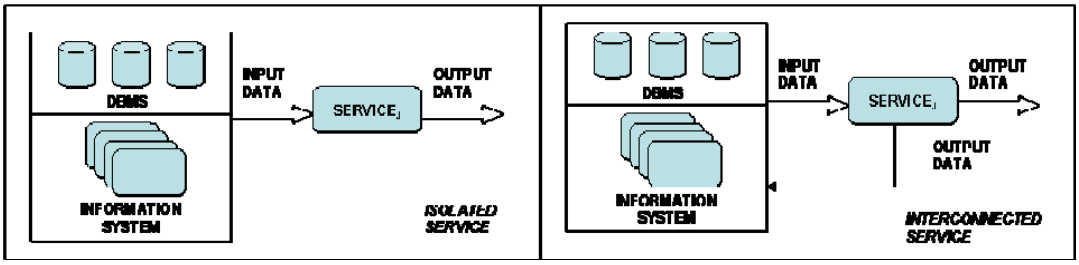

Figure 1 - Isolated and interconnected services

Systems that provide isolated services can be compared to open loop systems in which the improvements in the quality of ouput data are totally dedicated to meet the user requirements. In fact, they do not impact on the provider operational processes and the benefits 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 operational 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 operational processes. In this case, improvements decreases 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.
Furthermore, it is necessary to clarify that in general organizations offer different types of service to satisfy different user requirements. It is possible to define classes of users that access the same set of services with the same data quality requirements. The framework presented in this paper assumes that when a user belonging to a user class accesses a service, the system automatically associates the request with a specific set of quality requirements.

The personalization of services to user expectations is a particularly relevant research theme in the service oriented architecture design literature and it usually performed by applying *profiling* techniques. Profiling is the technique through which data are collected and manipulated with the goal of identifying and describing the profile of an entity, such as a user, an object, a product or a process [16]. A profile is a structured representation of the information that describes a user and his/her preferences along the services and the data that he needs to access. Generally, a user profile describes user requirements with a list of *<attribute-value>* pairs, where *value* describes a specific user or user class. Note that users can express a specific requirement or can be associated with a requirement only because they belong to a user class. In the first case the requirement is *explicit*, otherwise it is defined as *implicit*. The explicit requirements concerning data quality are expressed with the indication *<data quality dimension - value-weight>*, where *value* is the minimum level of acceptability specified by the user and *weight* is the importance of the dimension.
As regards implicit requirements, they are related to the processes that manipulate data and, consequently, to the services that the user accesses. Implicit requirements associated with users classes are important since users are rarely able to define their own quality requirements.

In short, our model of service offerings 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. First of all, users belonging to the same class are associated

with the same quality requirements along the same service S_j. We can define, for each service and for each user class, the data quality level $qc_{k,j}$ defined in the service offerings QC_j for service S_j. Note that each $qc_{k,j}$ 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 for service S_j is defined as a set of increasing data quality levels associated to K classes, that is:

$$\boldsymbol{QC_j} = \{\boldsymbol{qc_{1,j}}, \ldots, \boldsymbol{qc_{K,j}}\}. \quad (2)$$

The *data quality level* $qc_{k,j}$ defined for a user class can be distinguished from a *subjective quality level*, which can be specified by each user. In fact, users belonging to a specific class can have individual quality requirements that are different from their class requirements. In spite of simplicity, we assume that users are only associated with the class requirements. The extension of the model to the subjective model would not change the formalization of the model. In fact, in order to formalize a subjective quality level, it is possible to suppose that users inherits all the characteristics of their user class and can specialize inherited requirements by specifying his individual minimum quality levels.

From the providers perspective, the aim is to define service offerings QC_j that satisfy 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.

5. A UTILITY-BASED MODEL FOR DATA SERVICE PROVISIONING

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 a data service scenario. In defining the model, we refer to simple settings in which only one service S is provided. However, we argue that the model can be easily generalized to the case of the provider selling different kind of services.

The model relies on the definition of utility functions for both data providers and users. In our model, we adopt quasi-linear utility functions [9]. Quasi-linear utility functions represent an efficient and compact modeling tool for situations in which it easy to isolate, for every participant, positive utility terms (*value*) and negative utility terms (*payments*). We argue that the case of data quality and, specifically, data 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 analized by a large body of academic literature [3, 5, 8, 16, 21].

Quasi-linear utility functions are such that the utility value for an agent on a given contract 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, he/she sustains a cost for providing the negotiated contract, therefore losing value. 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 users in the multi-class data service scenario introduced in the previous section. The definition of utility functions is parameterized to model the two cases of isolated and interconnected services. Then, we show how the utility model can be exploited to provide a preliminary criterion for the provider to define optimal service offerings and, consequently, clarify which quality improvements need to be performed.

5.1 Data Providers Utility

Generally, a quasi-linear utility function defined for an agent P behaving in a service provider's perspective is defined as:

$$\boldsymbol{U_P(X) = P(X) - C_P(X)}; \quad (3)$$

where $P(X)$ is price, that is, the amount of money obtained by P for providing the generic contract X (payment term), whereas $C_P(X)$ represents the cost sustained by P to provide the contract X (negative value term).

In our multi-class data service scenario, the total amount of money 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 offering, that is:

$$\boldsymbol{P(QC) = \sum_{k=1}^{K} p(qc_k) \cdot M_k}, \quad (4)$$

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 $C_P(QC)$ represents the cost sustained by the provider to provide a service offerings QC. Such term keeps trace of two main types of quality costs. On the one hand, the provider sustains a cost for acquiring data, which can be, for instance, the cost of the people that input them in the system or the cost of the acquisition of external data. On the other hand, another type of cost is represented by the cost of the quality improvement of acquired data. The quality improvement must be performed when the provider needs to define service classes for which the quality level qc_k is greater than the quality level of the acquired data. We argue that $C_P(QC)$ is a function of only the maximum quality level defined in service offerings, that is, $C_P(QC)=C_P(qc_K)$. In case, in fact, the provider has managed to raise the quality of its data to qc_K, the provider does not sustain any further costs if a quality level qc_k, with $k<K$, needs to be defined.

As stated before, the quality improvement of data may also have a positive effect on the organization. In case of interconnected services, in fact, since the data acquired and provided to customers are exploited by the organizational operational processes of the provider, a quality improvement is likely to result in a monetary benefit $B(QC)$ for the provider. The benefit $B(QC)$ of quality improvement in the interconnected service scenario results from a variety of benefits, such as reduced cycle time for processing orders or reduced complaints from customers as a consequence of increased data accuracy. Anyway, note that one of the main direct benefits is the reduction of non quality costs. Although many authors in the data quality literature advocate the importance of organizational benefits of data quality improvement [7], the concrete evaluation of such benefits is still an open research issue [3]. Also in this case, we argue that benefits are a function of only the highest quality level qc_K that appears in the service offering QC, since the provider, for its internal processes, is likely to use data of the highest possible quality.

In our opinion, the benefits are influenced by the *degree of interconnection*, that is, the degree with which the data provided to customers are exploited in the provider's organizational processes. For example, if we consider the execution of a money transfer, data produced by this transaction are provided to the customers and are also used in the operational system of the organization. The degree of interconnection is low since output data impact only on the operational level. Conversely, if an organization uses the data of the customers transactions to analyze customers' behaviour and enable service personalization, the degree of interconnection is high: output data, in this case, will have an impact on the operational, decisional, and strategic organizational levels. Generally, we argue that benefits arise from data correctness and they are dependent on the use of data.

We model the difference between isolated and interconnected services by introducing the *state of interconnection* α of a data service. For a given provider and a given service, the coefficient α is a Boolean value, where $\alpha = 0$ represents the case of isolated services, whereas $\alpha = 1$ is the case of interconnected services. Hence, the provider utility $U_P(QC)$, for a service offering QC and for a given set of user classes can be defined as:

$$\boldsymbol{U_P(QC) = P(QC) + \alpha \cdot B_P(QC) - C_P(QC) = \sum_{k=1}^{K} P(qc_k) \cdot M_k - C_P(qc_K) + \alpha \cdot B_P(qc_K)} \qquad (5)$$

5.2 User Utility

Generally, a quasi-linear utility function defined for an agent C behaving in the customer (user) perspective is defined as:

$$\boldsymbol{U_C(X) = V_C(X) - P(X)} \qquad (6)$$

where $V_C(X)$ is the value term for the customer, that is, the monetary value that the customer obtains from the provisioning of the service, whereas $P(X)$ is the price paid for the obtaining the contract X.
In the data service scenario considered in this paper, we take the detail level of user classes UC_k and, therefore, we first define the utility associated to users assigned to a user class UC_k:

$$\boldsymbol{U_C(qc_k) = M_k \cdot V_C(qc_k) - M_k \cdot P(qc_k)} \qquad (7)$$

where $V_C(qc_k)$ is the value generated from the data quality level qc_k for a single customer in class UC_k, while $P(qc_k)$ is the price paid by a single customer in class UC_k. Also in this case, M_k represents the number of users in the k-th class.

As a consequence, the aggregate utility of customers can be defined as the sum of utilities associated to the total number of customers in each user class:

$$\boldsymbol{U_C = V_C(QC) - P(QC) = \sum_{k=1}^{K} V_C(qc_k) \cdot M_k - \sum_{k=1}^{K} P(qc_k) \cdot M_k} \qquad (8)$$

It has to be noticed that the customer utility does not include, besides a value and a payment term, any other terms. We hypothesize that all the benefits derived from data provisioning are taken into account by the value term $V_C(QC)$. Conversely to what happens for the data provider utility function, values and payment terms in $V_C(QC)$ depend on the whole service offerings $QC=\{qc_1,\ldots,qc_k\}$. Customers, in fact, define their utilities on the level of quality that they receive. Hence, the aggregate customers' utility must take into account all the quality levels qc_k defined in the service offerings.

5.3 Exploiting the Utility model

Once having defined utility functions for data providers and customers, we show how the utility model can be exploited for developing some informed considerations on the definition of optimal service offerings for the provider and for the identification of quality improvement activities. In order to do that, we refer to the basic principles of mechanism design in microeconomic theories [9].

In the context of mechanism design, such as, for instance, in optimal auction design [12], there are two main approaches to evaluate the properties of a mechanism, i.e., *utility maximization* and *allocation efficiency*. The utility maximization perspective usually takes the point of view of one participant in the allocation problem, i.e., either the provider or customers, and it pursues the objective of maximizing the utility of such participant. As already pointed out in the Introduction, the Zero Defect approach to data quality can be assimilated to the customer utility maximization approach. The main assumption of the Zero Defect approach is to associate the optimal level of quality of data provided by an organization to the one that completely satisfies the requirements expressed by the customers. In a utility theory perspective, the customers' utility clearly springs from the satisfaction of their requirements. Therefore, providing a full satisfaction of customers' requirements can be easily thought as a way to maximize the utility $U_C(QC)$ of customers of a given service offering QC. As for the Zero Defect approach, previous work in data quality focuses on the customers' requirements and it is usually not concerned, at least in a first approximation, with the interests of providers.

In this paper, by means of the previously introduced utility model, we tackle the problem of service offerings for data services definition by referring to the other perspective of mechanism design, i.e., the perspective of allocation efficiency (see Figure 2). The objective, in this case, is to jointly consider the interests of data providers and customers.

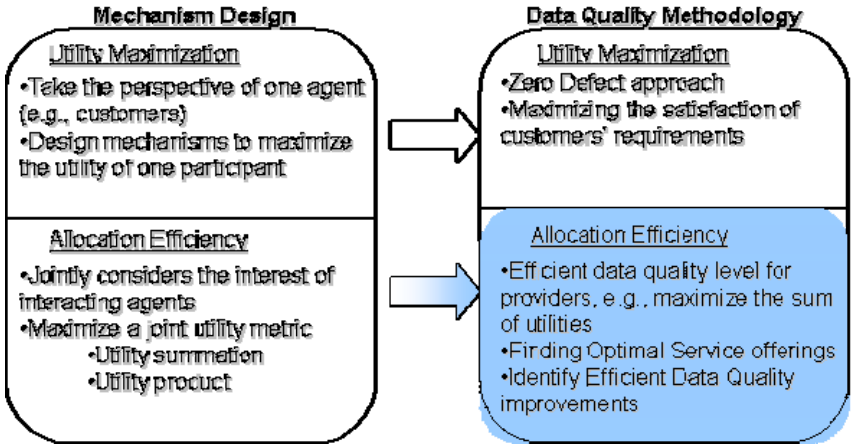

Figure 2 – The metaphor of mechanism design for data services contexts

According to this perspective, the best solution for the regulation of the data marketplace problem is to maximize a metric that jointly considers both the provider and the customers' utilities. The maximisation of the summation of the provider and customers utilities has been demonstrated, in auction theory, to represents a suitable way to jointly take into account the interests of all the participants in the mechanism. In this paper, we develop our criterion for defining optimal service offering on the basis of such metric for defining allocation efficiency. The literature on mechanism design, however, defines other metrics for allocation efficiency, such as the Nash equilibrium, identified by the maximization of the product of agents' utilities, which satisfies increasingly severe properties of fairness.

The allocation efficiency perspective is suitable to regulate long-term relationships between providers and customers. In the remainder, we show a preliminary consideration on the application of the allocation efficiency perspective in the context of data services.

Let us consider the allocation efficiency problem in the data service scenario in which we are to maximize the sum of utilities of the data provider and the customers.

With reference to Section 5.1 and 5.2, the sum U_P+U_C of the two aggregate utility values can be evaluated as:

$$\boldsymbol{U_P + U_C = \alpha \cdot B_P(QC) + V_C(QC) - C_P(QC)} \quad (9)$$

The function U_P+U_C is a function of the data quality offering QC implemented by the provider. Then, in order to identify the service offering QC* that maximizes the sum of the provider and the customers' utilities, we need to evaluate the derivative of the function U_P+U_C, w.r.t. to QC and put it equal to 0:

$$\frac{\boldsymbol{d(U_P + U_C)}}{\boldsymbol{dQC}} = \alpha \cdot \frac{\boldsymbol{dB_P}}{\boldsymbol{dQC}} + \frac{\boldsymbol{dV_C}}{\boldsymbol{dQC}} - \frac{\boldsymbol{dC_P}}{\boldsymbol{dQC}} = 0 \ . \quad (10)$$

Hence, the condition to be solved for finding the optimal service offerings QC* is as follows:

$$\alpha \cdot \frac{\boldsymbol{dB_P}}{\boldsymbol{dQC}} = \frac{\boldsymbol{dC_P}}{\boldsymbol{dQC}} - \frac{\boldsymbol{dV_C}}{\boldsymbol{dQC}} \ . \quad (11)$$

The aforementioned condition leads to some conclusions that extend the common assumptions for which providers should always adapt their service offerings to users requirements.

First, we want to stress that the configuration that maximizes the summation of utility of the provider and customers is the one for which the marginal benefit for the provider derived from savings in the costs of non-quality equals the difference between the marginal sustained costs and the marginal value created for the customers. In other words, the effort of the data provider should be devoted to find the quality improvement activities and the service offerings which achieve a perfect balance between the costs that must be sustained to perform the improvement and the value created for the customers. This assumption can be clarified by considering the two extremes of isolated and interconnected data services. For isolated services ($\alpha=0$), the maximization of the utility summation is achieved when there is a perfect balance between the marginal costs sustained by the provider to provide quality of data and the value created for the customers, that is:

$$\frac{\boldsymbol{dC_P}}{\boldsymbol{dQC}} = \frac{\boldsymbol{dV_C}}{\boldsymbol{dQC}}, \ \textit{for isolated data services.} \quad (12)$$

Consider for example a bank that accesses an external data source about stocks value in order to provide real-time information (thus to assure high timeliness) to its traders. The condition in Eq. 12 states that acquisition costs associated with the external data should be equal to or lower than the value of data for the interested customers.

At the same time, when the provider achieves certain benefits in its internal processes from the quality improvement ($\alpha=1$), the optimal service offerings is such that the marginal costs sustained by the providers are, at least partially, balanced by the benefits introduced by the quality improvement on

internal processes. Continuing with our example, the bank could spend more for data acquisition if the same data are used from internal bank operators to perform investment actions that would increase the bank profit.

Note that the utility model introduced so far leads to some considerations that differ from the common perceptions on data quality implied by classical data quality approaches. According to our perspective, the objective of the data provider is not to satisfy exhaustively the data quality requests of their customers, but it is rather the evaluation of an optimal set of data quality offering *QC**. In particular, the provider should privilege those quality improvement activities which create a marginal value for customers that is balanced by the difference between marginal costs sustained for the improvement and marginal benefits derived from the savings costs of non-quality on internal processes. The situations concerning too high improvement costs or too high marginal value for customers, which are not balanced by a return on internal processes should be avoided, since they do not lead to the maximization of the utility summation and, therefore, to an efficient allocation of the quality improvement efforts. Therefore, our model does not imply, in the provider perspective, the full satisfaction of the customers quality requirements. Our model should be used to identify a specific set of requirements expressed by customers that must be satisfied in order to achieve the maximization of the provider and customers' utility. Such requirements are the ones that, in order to be fulfilled, require a quality improvement effort which balances the marginal value created for customers, and, for the provider, the marginal costs sustained for the quality improvements and the benefits derived from savings in the non-quality costs.

Since we have made explicit (see Eq. 5 and 8) the relation between costs $C_P(QC)$, the customer value $V_C(QC)$ and the data quality offering QC, Eq. 11 can be used to evaluate the optimal data quality offering QC^* which results in an efficient allocation of quality improvement efforts of the provider, while maximizing the summation of the provider and the customers utility. Specifically, by considering the explicit relation between provider and customers' utilities and the definition of service offerings $QC=\{qc_1,\ldots,qc_k\}$, we can rewrite Eq. 11 as:

$$\alpha \cdot \frac{\boldsymbol{dB_P(QC)}}{\boldsymbol{dqc_K}} = \frac{\boldsymbol{dC_P(QC)}}{\boldsymbol{dqc_K}} - \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \frac{\boldsymbol{dV_C(qc_k)}}{\boldsymbol{dqc_k}} \qquad (13)$$

As stated in Eq. 13, *(i)* the marginal benefits and costs for the provider are dependent only on the maximum quality level qc_K in the service offering QC and *(ii)* the full structuring of the service offering has an impact on only the marginal benefits of customers. Therefore, we argue that the definition of the optimal data quality service offering QC for service providers should start with the assessment of the maximum quality level qc_K. Then, the intermediate levels qc_i, $i=1,\ldots,k-1$, need to be defined according to the condition specified in Eq. 13. In particular, in case of isolated services ($\alpha=0$), intermediate levels should be defined in order for customers' marginal values to meet the marginal costs savings defined by the maximum quality level qc_K. Similarly, in case of interconnected services ($\alpha=1$), the marginal value created by intermediate levels should match the difference between marginal costs and benefits derived from the choice of the maximum levels qc_K in QC.

6. CONCLUDING REMARKS AND FUTURE WORK

The objective of this paper has been to introduce a new perspective on the evaluation of the optimal of quality improvement efforts for data providers. The assumptions made by previous work imply the adoption of the zero defect approach that addresses the exhaustive fulfilment of the quality requirements expressed by customers. In this paper we have shown how, in order to maximize the summation of the

provider and the customers' utilities, the quality improvement activity and, consequently, the service offering, should privilege specific requirements.

We argue that the model presented in this paper can be used to set a new research agenda for the definition of optimal quality offerings for service providers in multi-class data service scenarios. However, we also need to stress at least two main limitations of the model, which set the stage for further developments of the work presented in this paper. First, although we provided a model for multi-class service provisioning, the exploitation of the utility model strongly relies on aggregated utility values. Future work should investigate more in depth how the definition of data service offerings impacts on the evaluation of the provider and the customer utility. Second, this paper bases the definition of optimal service offerings on the maximization of the provider and the customers' utility functions. Further considerations are likely to be introduced when the definition of optimal service offerings is made on the basis of other criteria for allocation efficiency, such as Nash equilibria. Finally, an open research issue concern the assessment of the provider and the customer interests on data quality. While our model also includes the definition of subjective quality levels for customers, the utility model remains defined on the basis of objective service classes set by the provider. Future work needs to investigate the impact of subjective utility assessment on the definition of optimal service offerings for the service provider.

REFERENCES

[1] Ballou D. P., Pazer H.L., Modelling Data and Process Quality in Multi-input, Multi-output Information Systems. *Management Science*, vol. 31, No. 2, February 1985.

[2] Ballou D. P., Wang R., Pazer H.L., Tayi G.K., Modelling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, vol. 44, No. 4, April 1998.

[3] Batini C., Cabitza F., Cappiello C., Francalanci C., A Comprehensive Data Quality Methodology for Web and Structured Data. In *Proceedings of the 1st International Conference on Digital Information Management (ICDIM '06)*, Bangalore, December 2006.

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

[5] Cappiello C., Ficiaro P., Pernici B. HIQM: a Methodology for Information Quality Monitoring, Measurement, and Improvement. Accepted for the publication in *Proceedings of the International Workshop on Quality of Information Systems* in conjunction with the 25th International conference on Conceptual Modeling (ER 2006)

[6] Cappiello C., Comuzzi M., Plebani P., On Automated Generation of Web Service Level Agreements", In *Proceedings of the International Conference on Advanced Information Systems Engineering (CAiSE'07)*, Trondheim, Norway, June 2007 .

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

[8] Martin J. Eppler, Markus Helfert: A Framework For The Classification Of Data Quality Costs And An Analysis Of Their Progression. *Proceedings of the International Conference on Information Quality (ICIQ'04)*, pp. 311-325

[9] Jackson, M.O., Mechanism Theory, In the *Encyclopedia of Life Support Systems*, edited by Ulrich Derigs, in the, EOLSS Publishers: Oxford UK, 2003

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

[11] Keller, A., Ludwig, H.: 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 (2002).

[12] Klemperer, P., Auction Theory: a Guide to the Literature. *Journal of Economic Surveys*, 13(3), pp. 227-286.

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

[14] A.Motro, P.Anokhin, and A.C. Acar. Utility-based resolution of data inconsistencies. In *Proceedings of the*

International Workshop on Information Quality in Information Systems 2004 (IQIS'04), Paris, France, June 2004.

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

[16] Olson, J. *Data Quality: the Accuracy Dimension*. Morgan Kaufmann, 2002.

[17] Orr K., Data Quality and Systems Theory. *Communications of the ACM*, vol.41, no.2, February 1998

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

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

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

[21] Wang R.Y., A Product Perspective on Total Data Quality Management. *Communications of the ACM*, vol. 41, no.2, February 1998.

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

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