UNDERSTANDING IMPARTIAL VERSUS UTILITY-DRIVEN QUALITY ASSESSMENTS IN LARGE DATASETS

(Completed Paper, IQ Measures)

Adir Even and G. Shankaranarayanan
Information Systems Department
Boston University School of Management
adir@bu.edu, gshankar@bu.edu

Abstract: Establishing and sustaining very high data quality in complex data environments is expensive and often practically impossible. Quantitative assessments of quality can provide important inputs for prioritizing improvement efforts. This study explores a methodology that evaluates both impartial and utility-driven assessments of data quality. Impartial assessments evaluate and measure the extent to which data is defective. Utility-driven assessments measure the extent to which the presence of quality defects degrades utility of that data, within a specific context of usage. The quality assessment methodology is empirically assessed using real-life alumni data – a large data resource that supports managing alumni relations and initiating pledge campaigns. The results provide important inputs that can direct the implementation and management of quality improvement policies in this data repository.

Key Words: Data Quality, TDQM, Information Products, Data Management, CRM

INTRODUCTION

Establishing and sustaining very high data quality is desirable from the data consumer's perspective. However, with rapidly increasing data volumes, this goal can rarely be achieved. Sustaining perfect quality is costly and often practically impossible. From an economic perspective, targeting perfect quality might be sub-optimal, as the cost of improving quality offsets the benefits gained. Given the difficulties in reaching perfect quality and the economic tradeoffs associated with sustaining it, there is a clear need to prioritize data quality improvement efforts and preferentially treat certain data elements or data subsets.

Quantitative assessments can provide important inputs to data quality management and direct improvement efforts and policies. Today, such assessments are mostly impartial, measuring the extent to which quality defects exist, disregarding usage context. In this study, we suggest that quality assessment can be significantly enhanced by considering the utility of data, the value contribution of data within a context of usage. We refer to this as utility-driven assessment of data quality. We develop a methodology that measures both impartial and utility-driven quality along different quality dimensions (illustrated here using completeness and currency). The results of this evaluation offer insights on quality characteristics and guide the development of quality improvement policies. We demonstrate this methodology in the context of Customer Relationship Management (CRM), using large samples from a data resource used by a university for managing alumni relations, soliciting donations, and initiating pledge campaigns.

The utility of information resources is derived from use [14], and depends on usage context (e.g., a decision task). The same data resource may have different utility in different contexts and, accordingly, the presence of data quality defects may differentially impact the degradation of utility. We therefore suggest that measuring quality as the extent to which utility degrades, affords contextual assessment of the impact of quality defects. Research in data quality has highlighted the importance of contextual assessment (e.g., [9], [12], [16]), but does not minimize the value of impartial quality assessments. Our objective here is to illustrate the application of the utility-driven assessment of data quality, using a real-life data environment, and highlight its implications for data quality management. This paper makes several important contributions. First, it validates the utility-driven assessment technique proposed in [5] by illustrating its application in a real-life data environment. Second, it offers a comparative analysis of impartial and contextual quality assessments and shows how the inter-relationships between these can offer important insights. Third, a utility-driven analysis of the data shows that individual dataset records may significantly differ in utility contribution. Further, different types of defects may affect the utility contribution of records differently, and this difference is reflected in utility-driven measurements. The study illustrates how such differences have important implications for managing data quality in large datasets; specifically, in terms of prioritizing quality improvement efforts. It is important to note that the variability in utility and its reflection in quality measurements are context-specific. Generalizing these results to other datasets (even within the same domain) requires repeating the evaluation. The final contribution here is the illustration of how this technique may be applied to understand utility and associated variations, and how such observations can guide the implementation of data quality improvement methods and policies.

In the remainder of this paper, we first discuss the challenges in managing the quality of large data resources and briefly review methods for assessing and improving quality that influence our research. We then propose a methodology for quality assessment, driven by the utility contribution of records. We apply this methodology to the alumni data and use the results to formulate quality improvement efforts that must be applied to this data resource. We finally highlight the contributions of this study, discuss managerial implications, and propose directions for further research.

BACKGROUND

High data quality is critical for successful integration of information systems in organizations. Data is subject to different types of quality defects – e.g., missing, corrupted, inaccurate, invalid or outdated content [5]. The presence of defects degrades quality, harms usability, and damages revenues and credibility [10]. Recent trends (e.g., data warehousing, enterprise resource planning (ERP), RFID, Supply-chains, and Clickstream) have mandated the need for complex data analysis. Consequently, organizations manage large and complex data resources. Targeting defect-free datasets in complex data management environments can be very expensive and often practically impossible. Further, targeting quality levels along multiple dimensions (e.g. accuracy vs. timeliness, completeness vs. consistency) can have inherent tradeoffs [1], [2]. Efficient quality management requires assessing these tradeoffs, optimizing (not necessarily maximizing) quality levels while allowing some imperfections [6], and prioritizing improvement efforts accordingly. Methods for data quality improvement can be classified into three general categories [10]:

a) Error Detection and Correction – errors can be detected by comparing data to a correct baseline (e.g., real-world entities, predefined rules/calculations, a value domain, or a validated dataset). Algorithms for automated detection/correction have been suggested (e.g., [8], [15]), and several commercial software packages support automated error detection and data cleansing [13]. When automated correction fails to achieve the desired results, firms may consider manual correction, or hiring external agencies that specialize in data cleansing. While error detection/correction can help raise the quality level to the desired target, it cannot fix root causes and prevent recurrence [10].

b) Process Control and Improvement – the Total Data Quality Management (TDQM) [17] suggests a continuous cycle of defining quality requirements, measuring along these definitions, analyzing the results and improving data processes accordingly. Unlike error detection and correction, this methodology can help detecting and fixing root causes and has shown to be successful in preventing recurrence. Different methodologies support the TDQM cycle – e.g., the Information Processing Map (IPMAP) for documentation [12], optimization of quality tradeoffs [3] and tools for visualization of quality trends [9].

c) Process Design – data processes can be built from the start (or existing processes redesigned) such that quality is more manageable and the likelihood of errors is smaller. Process design principles are discussed in a plethora of studies (e.g., [3], [10], [17]) – e.g., management involvement, embedded control, data modeling, processing procedures, and operational efficiency.

The methodology proposed offers important insights along all these categories. Specifically, this study addresses the prioritization of improvement efforts with respect to a large tabular dataset – multiple records with identical attribute structure. Although the method proposed can be applied to any tabular dataset, we focus on large tables in a data warehouse (DW), differentiating between two categories – *fact* and *dimension* (Figure 1). Fact tables capture transactional data. Depending on the database design, a record may represent a single transaction or an aggregation. A fact record includes numeric measurements (e.g., quantity and amount) and descriptors (e.g., time-stamps, payment/shipping instructions). It also includes dimension identifiers that link transactions to the business entities that describe them (e.g., customers, products, locations). The dimension table stores a list of dimension instances and associated descriptors (e.g., time-stamps, customer names, demographics, geographical locations, products, and categories).

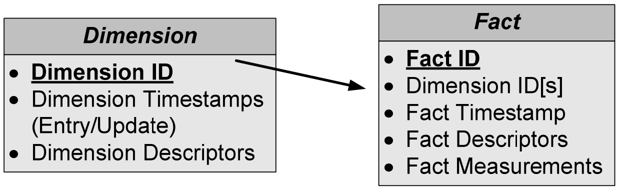

Figure 1: Dimension and Transaction Tables

This study addresses the quality improvement of dimensional data. Fact data, while not a subject for improvement in this study, is used for assessing the quality of dimensional data and developing improvement policies accordingly. The quality of dimensional data is critical in decision support environments. . For example - database marketing experts use sales data to analyze consumption behavior and manage promotion campaigns that target specific *customers* and *products* at specific *locations* [11]. Maintaining the associated dimensional data (i.e., customers, products, and locations) at a high quality is critical - otherwise, campaigns might fail to reach the right target. A common issue in data warehouses is the “slowly changing dimensions” [7] – dimension characteristics (e.g., income, marital status, and occupation) change over time and without proper tracking, transactional data and the associated dimensional data become unsynchronized and, hence, skew decisions.

Improving the quality of datasets has to consider: *(a) Target*: the targeted level can be evaluated along a continuum: at one end is a perfect quality level (i.e., data with no quality defects), and at the other end is a “hands off” approach - accepting quality as is, without making any efforts to improve it. Between these ends, we may consider a policy that improves quality to some extent but permits imperfections. *(b) Scope:*

we may consider an equal treatment of all records and attributes or, alternately, a differentiating policy – giving higher priority to improving the quality of certain records and/or attributes, and possibly making no significant efforts to improve others. Along these aspects, different types of policies can be evaluated:

Prevention: certain measures can be taken to prevent or reduce quality defects and the rate of their occurrence during data acquisition and processing - e.g., improving data acquisition interfaces, disallowing missing values, validations against a value domain, enforcing integrity constraints, or using a different (and possibly more expensive) data source with inherently cleaner data.

Auditing: quality defects may occur not only at acquisition, but also during data processing (e.g., due to miscalculation of new fields, or code-mismatch that incorrectly integrates multiple sources), or even after it has been stored (e.g., due to changes in the real-world entity that a dimension record describes). This requires auditing records, monitoring the process, and detecting the existence of defects.

Correction: even when defects are detected, correcting them is often questionable. In certain cases, correction is time consuming and costly (e.g., when a customer has to be contacted, or when missing content has to be purchased). One might choose to avoid the correction if the cost cannot be justified.

Usage: in certain cases, one might recommend users not to use certain subsets of records and/or attributes, or prevent usage altogether – e.g., when the quality is too low and cannot be significantly improved, or when the context of certain subsets turns out to be misleading in certain usage contexts.

Determining the target and the scope of certain policies has to consider the improvement that can be achieved, its impact on data usability, and the utility/cost tradeoffs that are associated with the implementation. Quantitative assessment of the anticipated utility/cost tradeoffs and the overall economic impact can help evaluate alternative policies and choose from among them [6]. The measurement methodology applied in this study can provide important inputs for such evaluation.

IMPARTIAL VERSUS UTILITY-DRIVEN ASSESSMENTS

Data quality is typically measured along multiple dimensions (e.g. accuracy, completeness, and currency), which reflect different hazards [16]. Quality is often measured on a scale between 0 (poor) and 1 (perfect) [9], [10]. Some methods are driven by physical characteristics (e.g., item counts, time tags, or failure rates) and assume an absolute and objective quality standard, disregarding the context in which the data is used. Alternative methods derive metrics from data content and evaluate them within specific usage contexts. The former approach is termed as structure-based (or structural), and the latter, content-based [2]. Quality can be measured impartially, representing perception that is based on the data itself regardless of usage, or contextual, reflecting usage-dependent perception [16]. In certain cases, the same dimension can be measured impartially and/or contextually, depending on the purpose of measurement [9]. As both impartial and contextual assessments contribute to the overall perception, it is important to address both. This study explores a methodology that evaluates the presence of quality defects (an impartial perspective) and their impact on utility degradation (a contextual perspective). Observing both perspectives is shown to provide important insights for quality improvement efforts and the development of associated policies [17].

This study adopts the measurement framework suggested in [5]. This framework, briefly described here, permits contextual measurement of quality along different dimensions and, with certain relaxations, allows impartial assessment as well. The quality measurement in this framework is driven by the utility of the dataset - a non negative measurement of its value contribution. This framework measures quality as a ratio; hence, it is indifferent to the utility-measurement units if used consistently. In this study, we consider the utility for a single usage; however, the framework in [5] accounts for multiple usages as well.

The evaluated dataset has N records (indexed by *[n]*), and M attributes (indexed by *[m]*). The data content of attribute *[m]* in record *[n]* is denoted $f_{n,m}$. The quality measure $q_{n,m}$ reflect the extent to which attribute *[m]* of record *[n]* suffers from a quality defect (between *0* - severe defects, and *1* - no defects). The overall utility U^D is attributed along records $\{U^R_n\}$, based on relative importance such that $U^D=\Sigma_{n=1..N}U^R_n$. The utility-mapping function u used in this framework links record contents and quality to its utility:

(1) $$U_n^R = u\left(\{f_{n,m}\}_{m=1..M}, \{q_{n,m}\}_{m=1..M}\right)$$

For a given set of attribute contents $\{f_{n,m}\}$, record utility reaches an upper limit U^{RMAX}_n when all attributes have perfect quality (i.e., $\{q_{n,m}=1\}$) and may be reduced by an extent when certain attributes are defective. The record quality Q^R_n is defined as a *[0,1]* ratio between the actual utility U^R_n and the upper limit U^{RMAX}_n:

(2) $$Q_n^R = U_n^R / U_n^{RMAX} = \left(u\left(\{f_{n,m}\}_{m=1..M}, \{q_{n,m}\}_{m=1..M}\right)\right) / \left(u\left(\{f_{n,m}\}_{m=1..M}, \{q_{n,m}=1\}_{m=1..M}\right)\right)$$

Similarly, dataset quality Q^D is the ratio between the actual and the maximum possible utility:

(3) $$Q^D = \left(\sum_{n=1..N} U_n^R\right) / \left(\sum_{n=1..N} U_n^{RMAX}\right) = \left(\sum_{n=1..N} U_n^{RMAX} Q_n^R\right) / \left(\sum_{n=1..N} U_n^{RMAX}\right)$$

When utility is allocated independent of attribute content (i.e., constant $U^{RMAX}_n=U^D/N$), the result is an impartial measure that reflects a ratio between the counts of perfect items and total items, which is consistent with common structural definitions (e.g., [9], [10]):

(4) $$Q_n^R = (1/M)\sum_{m=1..M} q_{n,m} \text{ , and } Q^D = (1/MN)\sum_{n=1..N}\sum_{m=1..M} q_{n,m}$$

This definition permits measurement along different dimensions, each reflecting a specific quality defect. For example, completeness reflects missing values, validity reflects failure to conform to a value-domain, accuracy reflects incorrect content, and currency reflects the extent to which data items are not up-to-date.

The magnitude of utility inequality is greater in some datasets than others [4]. The likelihood of the occurrence of quality defects in a record may be independent of its utility. However, recognizing a record as having a higher utility may encourage more focused efforts to reduce its quality defects. Utility-driven measurement reflects the impact of defects on the value contribution of the data, i.e., the extent to which utility is reduced by the presence of defects. Comparing the results of utility-driven to impartial assessments is important for managing quality in such datasets. At a high-level, we can differentiate between three cases with respect to such a comparison: *(a) Utility-driven scores are significantly higher than impartial scores:* this indicates that records with high utility are less defective. Two complementary explanations are possible: first, defective records are less usable to begin with, hence, have inherently lower utility. Second, some differentiating error-correction policies may have been applied – some efforts were made to maintain records with higher utility at a high quality level and eliminate their defects. *(b) Utility-driven scores not significantly different from impartial scores:* this indicates no association – the proportion of quality defects does not depend on the utility of certain records, whether high or low. This may also indicate high equality – utility that is nearly evenly distributed between all records, and *(c) Utility-driven scores significantly lower than impartial scores:* this indicates that records with high utility have a higher rate of quality defects. This abnormality may indicate a systematic cause of defects for record with high utility. This may also indicate high inequality in the dataset (i.e., a large proportion of utility associated with a small number of records), and some substantial damage to high-utility records. Understanding the relationships between impartial measurement and utility-driven measurement can help develop DQM policies, as demonstrated with our empirical assessment of the alumni data.

ASSESSING THE QUALITY OF ALUMNI DATA

To demonstrate utility-driven assessment of quality and its implications for prioritizing quality improvement efforts, we evaluate a sizably large sample of alumni data. This critical data resource helps generate a significant portion of the university’s revenue. The alumni data is used by different departments for contacting donors, tracking their gift history and managing pledge campaigns. This data resource, and the system that manages it, can be viewed as a form of Customer Relationship Management (CRM). Such systems are used for managing customer relations, tracking their past contributions, analyzing gifting patterns, and segmenting them for better targeting future promotion campaigns.

Methodology for Data Collection and Evaluation

This study uses samples from two key datasets, *Profiles* and *Gifts* (Table 1):

The *Profiles* dataset (dimensional data) has 358,372 records with contact and demographic data on alumni and other potential donors. The source dataset of the profile data has more than 100 attributes. Many of these are administrative, used for indexing and auditing purposes, hence, have low relevance to data consumers. In this study, we focus on six profile attributes that are extensively (based on our observations over the last year and interviews with key decision-makers) used for decision making: *School of graduation, Gender, Marital Status, Income, Ethnicity, and Religion*. These are all categorical attributes, i.e., each is associated with a value domain that consists of a finite set of possible values (stored in an associated lookup table). In addition, we observe two profile time-stamps: *Graduation Year*, in which a record was added to the dataset, and *Update Year*, in which a profile record was last updated.

The *Gifts* dataset (fact data) has 1,415,432 records and reflects the history of donations made. Some records in this dataset include data on actual gifts, while others contain data on commitments for future gifts (differentiated by a *Record Type* attribute). Each record has the gift amount and the year in which the gift was made, linked to a single profile.

Dataset	**Records**	**Growth**	**Attributes**	**Description**
Profiles – data on alumni, parents, and friends. One record per name listed	358,372	Annual average: 7,044 STD: 475	Profile ID	A unique identifier of the profile record
			Graduation Year	The year in which a profile record was added
			Update Year	The year in which a profile record was last updated
			School	The school from which the person graduated (28 categories)
			Gender	Male/Female
			Marital	Marital status (7 categories)
			Income	Income category (3 categories)
			Ethnicity	Ethnic group (7 categories)
			Religion	Religion (31 categories)
			Other Attributes	Contact information (e.g., address, phone), demographics, administrative fields
Gifts – detailed historical archive of gift transactions	1,415,432	Annual average: 45,884 STD: 6147	Gift ID	A unique identifier of the gift record
			Record Type	Some records represent pledges that have been paid later, or multiple payments on behalf of a gift
			Profile ID	A foreign key to the *Profiles* dataset. Each record is associated with one profile, but some profiles are not associated with any gifts.
			Gift Amount	The gift value (in USD)
			Gift Year	The year in which the gift record was added to the dataset
			Other Attributes	Additional details – e.g. pledge efforts, gift allocation, payment methods

Table 1: Alumni Data

Due to confidentiality constraints, our dataset includes only ~40% of the actual data volume, certain attributes were masked by codes, and gift amounts have been multiplied by a constant factor. The source data was collected between 1983 and 2006. In 1983 and 1984 (soon after initiation), a bulk of records that reflect prior activity were added (203,359 profile records, 405,969 gift records). Since 1985, both datasets have been updated regularly and steadily grown in size.

Our evaluation follows these steps:

(a) Preliminary evaluation: we collected summary statistics for all the variables used for quality assessment, and detected possible correlations and dependencies.

(b) Impartial quality assessment: we used the ratio measurements, which are based on item-counts (Equation 4), to evaluate impartial quality. Following [5], we initially considered four types of quality defects, with respect to the profile attributes that we evaluate:

1. *Missing values:* when recording a new profile (or updating an existing one) the source system permits leaving these attributes unfilled. A preliminary evaluation indicates that in a significant proportion of the records the values for these attributes are missing.
2. *Invalid data:* our initial evaluation indicated no invalid data with respect to the examined attributes, and all non-missing values conformed to the value domain.
3. *Up-to-date:* a significant number of profiles have not been updated for many years; hence, this is certainly a severe issue in this dataset. In some cases – they have never been updated since the record was added to the dataset. As a simple indicator of how current the record is, we use a binary variable – *1* if the record has been updated recently, and *0* if not. We evaluate this indicator both for a 1-year period (2006) and for a 5-year period (2002-2006). A more refined measurement is the exponential transformation [5] that converts age to a *[0,1]* measure:

   (5) $t = \exp\left\{-\alpha\left(Y^C - Y^U\right)\right\}$, where

   Y^C, Y^U - Current year (here, 2006), and the year of last record update, respectively

   α - A sensitivity factor, reflecting the rate of profiles becoming outdated. Here $\alpha=0.25$, assuming that *~20%-25%* of the profiles become outdated every year *($e^{-0.25}$ = ~0.77).*

   t - Up-to-date rank, ~0 for records that have not been updated for a long period (i.e., $Y^C >> Y^U$) and 1 for records that are up-to-date (i.e., $Y^C=Y^U=2006$).
4. *Inaccuracies:* according to the administrators of the source system, a significant number of profile records contain inaccurate attributes. This is mostly due to changes in a person's demographics that have been not tracked over the years, and less due to data-entry errors. However, due to the lack of appropriate baseline, in this study we could not evaluate the impact of inaccuracies.

Following this preliminary assessment, we focus on two defect types - missing values and up-to-date, and the associated quality measurements – completeness and currency, respectively. Completeness is evaluated at the data-item level (per attribute and overall), and at the record level. At the data-item level, impartial completeness is the ratio between the number of missing items and the total number. For assessing completeness at the record level, we consider two different methods: *(1) Absolute* – a record is marked as defective if at least one attribute (out of the 6 that are evaluated) has a missing value (i.e. 0 if defective, 1 if no defects are present), and *(2) Grade* – the number of non-defected attributes divided by the total number of attributes (i.e., a grade of 0 when all attributes are missing, 0.5 when half are missing, 1 where none are missing). It can be shown that calculating record-level completeness in the latter case is equivalent to calculating item-level completeness for all attributes combined. The last update time-stamp refers to the entire record and not to a specific attribute; hence, we have calculated currency at the record level only, using the binary indicators and the up-to-date rank.

(c)Utility-driven quality assessment: we repeat the quality assessment, using utility measurement as scaling factors (Equation 1-3). Using the *Gifts* dataset, we evaluate two utility measurements per profile:

1. *Inclination:* a binary variable that reflects a person's inclination to make a gift. This measurement has been evaluated for two time periods – the last 1 year (2006), and the previous 4 years (2002-2005). *21,485* profiles *(~6%)* are associated with donations in 2006, and *43,157 (~12%)* within 2002-2005.

2. *Amount:* the total amounts of gifts made; evaluated for the last 1 year and the previous 4 years.

These two utility measurements reflect different potential usages – *inclination*, for example, is likely to be observed for pledge campaigns that target a large donor base. *Amount*, on the other hand, is more useful for targeting specific donors who can potentially make very high contributions.

(d) Analysis: evaluating and comparing the results of impartial and utility driven quality assessments provides useful insights and has some important implications for developing DQM policies.

To demonstrate this calculation methodology, we use the illustrative sample of alumni profile data in Table 2, in which some attributes are missing (highlighted) and some records have not been updated recently.

ID	**Gender**	**Marital Status**	**Income Level**	**Record Complete (Absolute)**	**Record Complete (Grade)**	**Last Update**	**Recent Updated (1y)**	**Up-to-date Rank**	**Inclination**	**Amount**
A	Male	Married	Medium	1	1	2006	1	1	1	200
B	Female	Married	*NULL*	0	0.667	2003	0	0.47	1	800
C	*NULL*	Single	*NULL*	0	0.333	2005	0	0.78	0	0
D	*NULL*	*NULL*	*NULL*	0	0	1996	0	0.08	0	0
									2	***1000***

Table 2: Illustrative Alumni Profile Example

We observe that *2* out of *4* records are missing the value for *gender*; hence, impartial completeness with respect to this attribute is *0.5*. Similarly impartial completeness with respect to *marital status* is *0.75* (*1* out of *4* missing), and *0.25* with respect to *income level* (*3* out of *4* missing). For all attribute combined, the impartial completeness is *0.5* (*6* out of *12* missing). For record-level completeness – calculating along the *absolute* rank, 3 out of the 4 records have missing values (at least one attribute), hence, completeness is *0.25*. Using the *grade* rank, the record-level completeness (i.e., the average record grade) is *0.5*.

For utility-driven completeness measurement, we observe that only *2* out of the *4* profile records are associated with utility, and we use *inclination* and *amount* as scaling factors. With respect to *gender* and *marital status* – none of the utility-contributing records has missing values; hence, the utility-driven completeness is 1. With respect to *income level* – one utility-contributing record is missing the value. Factoring by inclination, the completeness is *(1*1+1*0)/2=0.5*, and factoring by *amount*, completeness is *(1*200+0*800)/1000=0.2*. At the record level, factoring the *absolute* rank by *inclination* yields a completeness level of *(1*1+0*1)/2=0.5*, and factoring by *amount* yields *(1*200+0*800)/1000=0.2*. Factoring the grade rank by inclination yields *(1*1+0*0.667)/1.667=0.6*, and by amount *(1*200+0.667*800)/1000=0.733.*

The impartial currency using the recent update indicator is *0.25*, and using the up-to-date rank is *0.58*. For utility-driven currency measurement, factoring by inclination yields *(1*1+1*0)/2=0.5* and *(1*1+1*0.47)/2=0.74*, respectively. Factoring by amount yields *(1*200+0*800)/1000=0.2* and *(1*200+0.47*800)/1000=0.58*, respectively.

Results

First, we have calculated the following variables for each profile record:

(a) Missing-Value Indicators: for each attribute (6 overall), the corresponding variable reflects whether the value is missing *(=0)* or not *(=1)*. We have also calculated for each record the absolute rank (*0* if at least one attribute is missing, *1* otherwise), and the grade rank (the average of the *6* attribute indicators)

(b) Up-to-date: we calculated a binary indicator that reflects whether a record has been updated within the last 1-year; and another for a 5-year period. We have also used the up-to-date rank, based on Equation 5.

(c)Utility Measurements: We have computed the *inclination* to donate (0 or 1) and the *total donation amount*, each for the last 1 year (2006) and the previous 4 years (2002-2005).

Summary statistics for these variables and the correlations between them are summarized in Table 3.

				Correlation*														
		Avg.	**STD.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Attribute	*1. School*	0.99	0.01	-	L*	L*	L*	L*	L*	L*	L*	L	L	L	L	L	L	L
	2. Gender	0.99	0.10	L*	-	M*	M*	M*	L*	L*	M*	L*	L*	L*	L*	L*	L	L
	3. Marital	0.89	0.30	L*	M*	-	M*	M*	M*	M*	H*	L*	L*	L*	L*	L*	L	L*
	4. Income	0.63	0.48	L*	M*	M*	-	M*	M*	H*	H*	L*	L*	L*	M*	M*	L*	L*
	5. Ethnicity	0.59	0.49	L*	M*	M*	M*	-	M*	H*	H*	M*	M*	M*	L*	L*	L	L
	6. Religion	0.60	0.49	L*	L*	M*	M*	M*	-	H*	H*	L*	L*	L	L*	L*	L	L*
Record	*7. Absolute*	0.36	0.48	L*	L*	M*	H*	H*	H*	-	H*	L*	L*	L*	L*	M*	L	L
	8. Grade	0.78	0.20	L*	M*	H*	H*	H*	H*	H*	-	L*	L*	L*	L*	M*	L*	L*
	9. Recent-1	0.17	0.37	L	L*	L*	L*	M*	L*	L*	L*	-	M*	H*	L*	L*	L*	L*
	10. Recent-5	0.51	0.50	L	L*	L*	L*	M*	L*	L*	L*	M*	-	H*	M*	L*	L*	L*
	11. Up-to-date	0.42	0.35	L	L*	L*	L*	M*	L*	L*	L*	H*	H*	-	L*	L*	L*	L*
Utility	*12. Inclin.1*	0.06	0.24	L	L*	L*	M*	L*	L*	L*	L*	L*	M*	L*	-	H*	L*	L*
	13. Inclin.2/5	0.12	0.33	L	L*	L*	M*	L*	L*	M*	M*	L*	L*	L*	H*	-	L*	L*
	14. Amt.-1	50	7.1K	L	L	L	L*	L	L	L	L*	L*	L*	L*	L*	L*	-	L*
	15. Amt.-2/5	190	11.7K	L	L	L*	L*	L	L*	L	L*	L*	L*	L*	L*	L*	L*	-

H: >0.5, M: 0.1 to 0.5, L: <0.1, (*) Significant (P-value < 0.02)

Table 3: Profile Variables – Summary Statistics and Correlations

Some insights from these results are summarized below:

- The rate of missing values is relatively high: *64%* of the records have at least one attribute missing (*Absolute* rank) and in average, *22%* of the attribute values are missing (*Grade* rank).
- A significant proportion of profile records are not up-to-date. Only *17%* of the profiles have been updated (or added) in the last year (*Recent-1*), and *49%* have not been updated in *5* years (*Recent-5*).
- The number of missing-values varies significantly between attributes – *School and Gender* have almost no missing values *(~0%*) while *Income*, *Ethnicity*, and *Religion* have a high rate (*~40%*).
- Correlations among missing-value indicators are mostly medium but significant. This implies that when a record is missing the value for one attribute, it is likely to miss values for other attributes as well.
- In most cases, missing value indicators have low, but significant, correlation with the up-to-date variables (indicators and rank). This implies that older records are somewhat more likely than newer records to have missing values, but not to a great extent.
- There is high and significant correlation between the inclination to donate and the total amounts in the most recent year, and the inclination/amount in the previous 4.
- Higher inclination to donate has high correlation with most quality indicators. Higher amount has significant correlation with all up-to-date indicators and with some missing-value indicators.

The last point suggests that higher impartial quality (less defects, more recent updates) is associated with higher utility. We next measure the extent of this association. For binary indicators (missing values, recent updates) we used a 2-way ANOVA test that measures the significance of the difference in utility between defective and non-defective records. For variables that are measured over a range (i.e., *Grade* and *Datedness* ranks), we use a linear regression. The P-values for these tests are summarized in Table 4.

	Variable	Inclination (1 Year)	Inclination (2-5 Years)	Amount (1 Year)	Amount (2-5 Years)
Attribute	*School*	0.620	0.945	0.972	0.939
	Gender	~0**	~0**	0.759	0.393
	Marital	~0**	~0**	0.207	0.008**
	Income	~0**	~0**	0.021*	~0**
	Ethnicity	~0**	~0**	0.598	0.048*
	Religion	~0**	~0**	0.060*	0.004**
Record	*Absolute*	~0**	~0**	0.067*	0.486
	Grade	~0**	~0**	0.029*	0.007**
	Recent-1	~0**	~0**	~0**	~0**
	Recent-5	~0**	~0**	0.001**	~0**
	Up-to-date	~0**	~0**	~0**	~0**

(**) Highly Significant (P-value < 0.01), (*) Marginally Significant (P-Value < 0.1)

Table 4: Significance of Utility Variance *(P-Value*)*

The results indicate that inclination to donate (for both periods) has a significantly strong association with almost all impartial indicators. The amount, on the other hand, has a significantly strong association with the up-to-date indicators, but only with *some* missing-value indicators. Importantly, the adjusted R-SQR results are low (below 0.1), implying that the variability in utility is associated with, but cannot be entirely explained by, high impartial data quality. We next measured impartial and utility-driven completeness and currency, following the methodology that was described earlier. The results are summarized in Table 5.

		Impartial Completeness	Utility-Driven Completeness			
			Inclination (1y)	Inclination (2-5 Y)	Amount (1 Y)	Amount (2-5 Y)
Attributes Completeness	*School*	0.999	0.999	0.999	0.999	0.999
	Gender	0.990	0.997	0.998	0.997	0.999
	Marital	0.894	0.950	0.958	0.984	0.977
	Income	0.631	0.872	0.896	0.891	0.836
	Ethnicity	0.596	0.646	0.654	0.656	0.496
	Religion	0.605	0.717	0.715	0.819	0.751
	All	0.786	0.863	0.870	0.891	0.843
Record	*Absolute*	0.356	0.497	0.511	0.561	0.608
Completeness	*Grade*	0.786	0.863	0.870	0.891	0.843
Record Currency	*Recent-1*	0.171	0.282	0.219	0.635	0.552
	Recent-5	0.510	0.635	0.635	0.899	0.860
	Up-to-date	0.425	0.540	0.518	0.807	0.775

Table 5: Quality Assessment

Most utility-driven data quality measurement scores are higher than their corresponding impartial measurement scores. This is not surprising since, along most indicators, higher utility has a significantly stronger association with higher impartial quality. However, some insights can be gained by observing the extent to which utility-driven measurements are higher and more consistent:

- Utility-driven completeness measurements, at the attribute level and at the record level, are relatively consistent along the four utility metrics. This implies that, when assessing the completeness of this alumni profile data, calculating utility-driven measurements along multiple utility metrics does not grant a significant advantage over measuring it along a single metric.
- For attributes with inherently high impartial completeness (e.g., *School* and *Gender*), utility-driven measurements are not substantially different from the impartial measurements. Some margin exists for *Marital Status* – but since the impartial completeness is relatively high, this margin is fairly small.
- For attributes with inherently low impartial quality, we see substantial differences in the margin between the impartial and the utility-driven scores. In the case of *Ethnicity*, the margin is relatively minor. It is slightly higher for *Religion*, and a lot higher for *Income*. This implies that these attributes have very different association with the utility gained. The completeness of *Income* data significantly differentiates between low-utility and high-utility profile records (both along *Inclination* and *Amount*). The completeness of *Religion* data also differentiates these, but to a lesser extent, and the completeness of *Ethnicity* does not significantly differentiate the profile records.
- Measuring completeness for all the attributes combined, or measuring it at the record level, has an averaging effect. Some margins exist between impartial and utility-driven measurements, but they are not as significant as the margins for the measurements associated with specific attributes.
- Unlike completeness, with respect to currency, amount-driven scores are significantly higher than inclination-driven scores along all indicators. This implies that the extent to which a record is up-to-date is significantly associated with the amount donated, beyond just the fact that a person has made a donation. This finding may suggest that currently the practice is to audit and update more frequently data on donors who have contributed or have a high contribution potential (as confirmed with the alumni data administrators). Notably, compared to the average, the variance of donation amounts is very large (in Table 3). This may suggest a significantly uneven distribution among the gift amounts that are associated with each profile – a small number of profiles are associated with large gift amounts, while a large number of profiles are associated with small gift amounts or with no gifts at all.
- With respect to utility-driven currency measurement, there is a significant difference between using *inclination* versus using *amount* as utility factors. However, there is no significant difference between evaluating utility (using *inclination* or *amount* as factors) over 1 year versus the previous 4. This can be explained by the high correlation between the measurements over different time periods.

Discussion

The results indicate association between utility and quality, with respect to profile data. Profiles that are up-to-date and missing fewer attribute values are associated with higher utility contribution, either when measured based on *inclination* to donate or by the total donation *amount*. Accordingly, utility-driven measurements are higher (significantly so in some cases) than impartial measurements. Based on our discussion with the data administrators, this association between quality and utility can be explained by:

- New profiles are typically imported from the student registration system, which can provide only a subset of the required attributes (e.g., *Income* level is not provided, *Ethnicity* and *Religion* are only partially available). As a result, most profile records enter the system with missing attributes, which negatively affects the ability to assess their potential contribution.
- Some profile attributes are likely to change over time (e.g., *Address*, *Telephone, Income, and Marital Status*). Failure to keep profiles up-to-date significantly limits the ability to contact the alumni, gather additional data, and assess their contribution potential.
- Data administrators and system users tend to update profile information and fill-in missing values only when a person makes a donation (e.g., by contacting the person and running a quick phone survey). As a result, if a person made a donation recently, his/her profile data is likely to be up-to-date and have less missing values. On the other hand, if a person has not made any donation for a few

years in a row, the quality of his/her profile is likely to deteriorate.

- In some cases, the data administrators and/or users update and enhance the data on certain donors by paying agencies that specialize in enhancing such data. However, this is mostly done for a limited number of donors who exhibit a high potential for future contributions. As a result, data in profile records associated with donations is likely to be maintained at a significantly higher quality level.

While the link between utility and quality is acknowledged by the administrators and key decision-makers that use this alumni data, and reflected to some extent in current data management policies, the results of our evaluation shed light on some issues that need further attention. The same results can also guide the development of better quality management policies for this data resource:

Differentiation: In general, the data administrators should clearly consider a differentiating policy with respect to auditing records and attributes, correcting quality defects, and implementing procedures to prevent defects from reoccurring. They may also consider recommending that data users refrain from using certain records or attributes for certain types of usages (decision tasks and applications). Our results indicate a significant variation in utility contribution among profile records. They also point to significantly different utility associations of the different data attributes. Utility appears to have a high sensitivity to the currency of updates and, lastly, the quality measurements along different quality dimensions are also different. With such extensive variations, treating all records and attributes identically is likely to be economically sub-optimal. Data quality management efforts and policies (e.g., prevention, auditing, correction, and usage) must be differentially applied to subsets of records in a manner that is likely to provide the highest improvement in utility for the investment (i.e., gaining the “biggest bang for the buck”).

Attributing Utility: Our results highlight the benefit of measuring and attributing utility. Our metrics, inclination and amount, reflect the impact of quality defects on utility; hence, permit convenient calculation of utility-driven measurements. Interestingly, for both metrics there was no significant difference between the quality scores for the two time periods (i.e., last 1 year versus previous 4 years). This can be possibly explained by the high correlation between donation patterns over time – a person who donates in a certain year is likely to donate also in the year after. Based on this observation, an important refinement to utility measurement is to consider not only past donation behavior, but also some prediction of the potential for future donations, e.g. by applying Customer Lifetime Value (CLV) measurement techniques [11].

Improving Completeness: The results indicate that analyzing the impact of missing values at the record level alone is insufficient. There is certainly a need to further assess the impact of missing values at the attribute level. The impartial completeness of certain attributes is inherently high (e.g., *School* and *Gender*, with nearly 0 missing values); hence, the potential to gain utility by correcting these attributes is negligible. Even for attributes with lower impartial completeness, we can expect substantial variability – with some attribute (e.g., *Income*) we may see a strong association between missing values and utility contribution. Such attributes obviously need to get a very high priority in terms of improvement efforts. With other attributes (e.g., *Marital Status* and *Religion*), we may see some association, but to a lesser extent. With yet other attributes (e.g., *Ethnicity*), the association, if at all, is very small. In the latter case, we may reconsider whether or not is it worthwhile to invest in any quality improvement efforts, or even consider giving up the storage and management of this attribute. Notably, the data resource evaluated here contains many (over a hundred) other profile attributes, and managing these could benefit from a similar evaluation.

Improving Currency: Utility was strongly linked to currency – outdated profiles are associated with lower inclination and amount. This indicates a need to audit profiles more often. Currently, approximately half of the profiles have not been updated within the last 5 years. The potential for contributing to utility can help prioritizing the update efforts. As shown earlier, there is a strong association between recent donations (last 1 year) and past donations (previous 4 years); hence, profiles that are associated with recent inclination to donate should be high priority for quality improvement efforts (e.g., a more frequent auditing). Another direction to explore is the ability to link inclination to donate to certain attributes – e.g. the *Income*. Such an attribute can serve as a classification category for setting up the update priorities – e.g., audit and update profile records associated with high income more often. Once an attribute has been selected as a classifier, its quality should be maintained at a high-level. For example, if *Income* is found to be a good predictor of utility, efforts should be made to keep it up-to-date and eliminate its missing values. We may also consider refining its granularity (currently, only 3 income categories are used), and adding a time stamp that tracks specific changes (currently, changes are tracked only at the record level).

The quality and the utility of alumni data certainly have room for improvement as only a relatively small number of profiles are associated with donations, and quality defects are present in high proportions. Importantly, our analyses do not offer a comprehensive solution for prioritization and policies, but rather demonstrates the methodology and the insights that one stand to gain from such analyses. A more complete solution demands an analysis of all relevant attributes, evaluation of other utility measurements, statistical tools for estimation of future benefits, and possibly a revision of existing data usage patterns.

CONCLUSIONS

Quantitative quality assessment is important for continuous improvement of data quality. Common measurement methods largely reflect an impartial perspective and disregard the context in which the data is used. This study explores a measurement methodology that reflects a contextual perspective as well, by observing not only the presence of defects, but also their impact on the utility gained. Applying both impartial and utility-driven assessments provides important insights on the strengths and weaknesses of current data quality management practices. It can direct the improvement of these practices and the development of new policies. The application of this methodology is demonstrated in the context of managing alumni data, showing how current quality measurement methods compare and are supplemented by the proposed method for measuring and improving data quality.

The results highlight the importance of understanding and assessing the utility of data resources. Different elements in a dataset (e.g., records and/or attributes) may significantly vary in their contribution to utility. In certain cases, a major proportion of utility may be contributed by a small subset of these elements, while in other cases the utility is distributed more evenly. Modeling and quantifying utility distribution and detecting possible inequalities can direct quality improvement efforts and help prioritize them. Utility assessment is also important in the presence of significant economic tradeoffs – certain improvement efforts are expensive, and their cost might offset the added utility. Evaluating both utility and cost along the same monetary scale can help assess these tradeoffs and detect economically-optimal policies.

This study is not without limitations. It evaluates the quality of a single tabular dataset (a dimension table in a data warehouse). Data management environments include multiple datasets and some of these may use non-tabular data structures. Further, the quality of transactional data needs to be improved as well. The utility measurements used here – donation inclination and amount – are specific to the customer relationship management domain. Other application domains and business environments (e.g., finance, healthcare, insurance), will require the identification of fundamentally different utility measurements that are specific to each. The study has evaluated utility for relatively recent usage. However, in almost all business settings, it is important to consider the potential future utility gain and develop quantitative tools for estimating it. This is particularly important for evaluating the quality of a new data source which has not been used before, or enhancing an existing data source with additional records and attributes.

The study examines two types of quality defects – missing values, reflecting the completeness dimension, and up-to-date, reflecting currency. Validity (or the lack of), reflecting data items that do not confirm to a value domain, is relatively easy to detect, and can be measured using the suggested methodology. It was found to be a non-factor in this alumni dataset, but can be a serious hazard in environments that integrate data from multiple sources. Inaccuracy, the presence of incorrect values, is a serious hazard in many data management environment, including the alumni data. Our measurement methodology can be applied for assessing accuracy as well. However, detecting incorrect values and fixing them can be significantly more difficult, as a baseline for comparison is not always available or easy to determine. Validating the accuracy of all records and attributes is pricy or practically impossible with a large dataset; hence, addressing accuracy will require the development of innovative statistical sampling methods.

Finally, our evaluation highlights causality in the relationships between utility and quality. Common perceptions see quality as antecedent to utility – reducing defect rate and improving quality level increases the usability of data, hence, the utility gained. Our results suggest that in certain settings a reverse causality may exist – frequent usage and high utility promote quality improvement of certain data elements, while the quality of items that are not frequently used (e.g., profile records associated with donors that have not donated in a long period) is likely to degrade. This mutual dependency may have positive implications (e.g., cost-effective quality improvement, as efforts focus on items that contribute higher utility), as well as negative (e.g., usage stagnation, a failure to realize utility potential of less-frequently used items due to degrading quality). This causality and its implications should be further explored and understood.

REFERENCES

[1] Ballou, D. P., and Pazer, H. L. "Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff." *Information Systems Research*, 6 (1). 1995. pp. 51-72

[2] Ballou, D.P. and Pazer, H.L. "Modeling Completeness versus Consistency Tradeoffs in Information Decision Systems." *IEEE Transactions on Knowledge and Data Engineering*, 15 (1). 2003. pp. 240-243.

[3] Ballou D. P., Wang R., Pazer H., and Tayi G. K. "Modeling Information Manufacturing Systems to Determine Information Product Quality." *Management Science*, 44 (4). 1998. pp. 462-484

[4] Even, A., and Shankaranarayanan, G. "Utility Inequality and Its Implications for Data Management." *The International Conference of Information Quality (ICIQ)*. Nov. 2006. Cambridge, MA

[5] Even, A., and Shankaranarayanan, G. "Assessing Data Quality: a Value-Driven Approach." *The DATA BASE for Advances in Information Systems*, 38 (2). 2007. pp. 76-93

[6] Even, A., Shankaranarayanan, G. and Berger, P. D. "Economics-Driven Data Management: An Application to the Design of Tabular Datasets." *IEEE Transactions on Knowledge and Data Engineering*, 19 (6). 2007, pp. 818-831

[7] Kimball R., Reeves L., Ross M., and Thornthwaite W. *The Data Warehouse Lifecycle Toolkit*, Wiley Computer Publishing, New York, NY, 2000

[8] Lee, Y.W., Pipino, L., Strong, D.M., and Wang, R.Y. "Process-Embedded Data Integrity." *Journal of Database Management* 15 (1). 2004. pp. 87-103.

[9] Pipino L.L, Yang, W.L. and Wang, R.Y. "Data Quality Assessment/" *Communications of the ACM* 45 (4). 2002. pp 211-218

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

[11] Roberts, M. L., and Berger, P. D., *Direct Marketing Management*, Prentice-Hall, Englewood, NJ, 1999

[12] Shankaranarayanan, G., Ziad, M., and Wang, R. Y. "Managing Data Quality in Dynamic Decision Making Environments: An Information Product Approach." *J. of Database Management* 14 (4). 2003. pp. 14-32

[13] Shankaranarayanan, G. and Even, A. "Managing Metadata in Data Warehouses: Pitfalls and Possibilities." *Communications of the AIS* 14(13). 2004. pp. 247-274

[14] Shapiro C., and Varian H.R. *Information Rules,* Harvard Business School Press, Cambridge, MA, 1999

[15] Tayi G.K., and Ballou D.P. "An Integrated Production-Inventory Model with Reprocessing and Inspection." *International Journal of Production Research* 26 (8). 1988. pp. 1299-1315

[16] Wang, R.Y. and Strong, D.M. “Beyond Accuracy: What Data Quality Means to Data Consumers.” *Journal of Management Information Systems* 12 (4). 1996. pp. 5-34

[17] Wang R.Y. “A Product Perspective on Total Quality Management.” *Communications of the ACM* 41(2). 1998. pp. 58-65