VALUE-DRIVEN DATA QUALITY ASSESSMENT

(Research Paper - IQ Concepts, Tools, Metrics, Measures, Models, and Methodologies)

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

Abstract: Techniques for assessing data quality along different dimensions have been discussed in the data quality management (DQM) literature. In recent years, researchers and practitioners have underscored the importance of contextual quality assessment, highlighting its contribution to decision-making. The current data quality measurement methods, however, are often derived from impartial data and system characteristics, disconnected from the business and decision-making context. This paper suggests that with the increased attention to the contextual aspects, there is a need to revise current data quality measurement methods and consider alternatives that better reflect contextual evaluation. As a step in this direction, this study develops content-based measurement methods for commonly-used quality dimensions: completeness, validity, accuracy, and currency. The measurements are based on *Intrinsic Value*, a conceptual measure of the business value that is associated with the evaluated data. Intrinsic value is used as a scaling factor that allows aggregation of quality measurements from the single data item to higher-level data collections. The proposed value-based quality measurement models are illustrated with a few examples and their implications for data management research and practice discussed.

Keywords: Data Quality, Data Quality Management, Information Products, Database, Metadata, Information Value, Decision Making

INTRODUCTION

Data has been recognized as an essential resource that supports a plethora of business activities. Resource investments and managerial efforts toward data management activities and related systems are steadily increasing [17, 32]. Traditional data management concepts and design methodologies are geared primarily towards ensuring functionality and technical efficiency. However, with the growing investments in data management, there is an increasing concern about the economic value of data and the return on investments in data management. This shift in focus has significant implications for data quality management (DQM). Data quality is a critical issue in information systems due to the rapid growth of data volumes and their complexity. The potential for capital losses and heightened risk exposure due to poor data quality makes data quality management critical in organizations [11, 32]. From a broader perspective, the quality of the information and of the systems that provide it have been identified as having a key influence on IS adoption and end-user satisfaction at the individual level, resulting in the positive contribution of the information system to organizational performance [9].

Academic research has addressed data quality management in a variety of different ways. Established total quality management (TQM) techniques from manufacturing have been extended to manage data quality (Total Data Quality Management or TDQM) [31]. Methods, that treat data as a product, have been proposed to model the data manufacturing processes and evaluate the quality of data [3, 5, 13, 21, 22]. Techniques that capture and use metadata to manage and improve data quality have also been described [18, 20, 26]. Statistical techniques to detect and correct data errors [23, 34], techniques that attempt to improve/define data quality using data source calculus and algebra [21], data stewardship [13], and dimensional gap analysis [19] have all been described in literature. Such techniques clearly contribute to better data and data quality management but often do not take into account important contextual factors such as the task for which the data is to be used, or the individual characteristics of the decision-maker. Contextual factors have been shown to strongly influence perceptions of data quality [18, 27], and researchers in the field are beginning to take contextual factors and individual differences into account when managing data quality [34].

While high quality data and associated data quality initiatives do benefit the organization, can this benefit be assessed from an economic perspective? Can we quantify the contribution of data quality improvement to decisions and actions? Do these benefits offset and supersede the implementation costs? Questions related to the economics of data quality, such as cost savings, revenue generation, and profitability, are challenging and not well researched. However, such questions are very important to business firms, especially with the high costs and efforts associated with quality improvement initiatives. Arguably, business value is not created by stand-alone data, but rather through the integration of the data into business processes and its use by decision makers [8]. Hence, the perception of quality will be influenced by contextual factors, such as the organizational level at which the data is used, the specific task, and/or the personal preferences of the decision maker. Contribution to business value is an important aspect of data quality that must be observed through a contextual lens. The same data may be valued differently in the context of different decision tasks.

This study suggests that acknowledging the importance of the contextual perspective implies a need to re-think current data quality assessment methods. Quantitative evaluation of data quality attributes such as accuracy, completeness, or currency, is often based upon impartial characteristics of the data and/or the systems that manage it – ratios between item counts, time measurements, or failure rates [16, 22, 23]. Ballou and Pazer [4] identify this approach as *structural* – quality measurement with an underlying assumption of absolute standard, disconnected from a specific usage. Alternatively, measurement can be derived by the *content* and its applicability for business use. Adopting this perspective, the research described here proposes measurement methods that link impartial characteristics and contextual perception by rooting the data quality assessment on *intrinsic value* - a conceptual measure of the business value associated with the data. A key contribution of this study is a quantitative integration of the impartial and the contextual aspects of data quality, towards a more useful and meaningful quality assessment. Value-based measurements can be valuable from different perspectives – individual, business, and IT-administration. For an individual data–consumer, value-based measurement defines a better link between the data quality measurements provided, and the contextual perception of quality associated with the decision task. For business management, it offers better insights on the economic assessment of data management systems and quality improvement initiatives. From the IT-administration perspective, it allows better prioritization of data quality management efforts.

The remainder of this paper is organized as follows: section 2 provides the theoretical background by reviewing data quality attributes and measurement methods that have been discussed in literature. Section 3 introduces the concept of intrinsic value related to data at different hierarchical levels. Quality measurement methods that use the intrinsic value as a scaling factor are introduced in section 4. The methods are developed for four commonly-used quality attributes – completeness, validity, accuracy and currency. Finally, section 5 discusses the implications of value-based quality measurement for IS research and practice, offers concluding remarks, and proposes directions for future research.

ASSESSING DATA QUALITY – RELEVANT LITERATURE

Data quality has long been a critical part of information system management and DQM literature has defined and characterized different perspectives of data quality and its management [29]. A consensus is that quality should be defined from the viewpoint of the data consumers and hence the fundamental definition of quality as "fitness for use" [23, 24, 30]. However, empirical studies have also shown that data quality is perceived as a complex and multi-dimensional concept [30]. Many studies have discussed the different quality dimensions, or attributes, such as accuracy, completeness, and consistency [16, 22, 23]. These attributes have also been conceptualized as representing the data consumer's perception of quality [28]. Furthermore, studies have suggested high-level categorization of quality attributes to reflect different aspects of functionality and design of information systems (e.g., infrastructure, model, process, contents, and presentation) [23, 30].

Quantitative assessment of the quality attributes has been identified as a key factor to successful data quality management [22, 23]. It has become an integral part of the Total Data Quality Management (TDQM), an adaptation of the TQM to data quality management based on the information product approach [31]. This approach perceives the output of a data management system as an Information Product (IP) that is aimed to satisfy the needs of the information customers. The collection of systems and processes that gather and transform the data and make it available to customers is viewed as a data manufacturing process (DMP), also referred to as the information value chain [23]. TDQM views data quality management as a continuous cycle of defining, measuring, analyzing and improving of the data manufacturing process to achieve the goal of continuously improving the end-product[1]. The outcome of the measurement stage, a set of data quality measurements along the different attributes, serves as the input for the analysis of the DMP for identifying problematic configurations, detecting root-causes of quality failures, and evaluating alternative policies for quality improvement. The contribution of quality measurements to the DMP management can be enhanced by capturing and storing them in a form of quality metadata [7, 26], providing dedicated software tools for calculation and presentation of the measurements [22, 25, 31], and linking the measurements to a process map of the DMP [24]. When used as an input to utility and cost functions, such quality measurements can serve to optimize the data manufacturing processes [5]. DQM studies have specifically addressed the RDBMS (Relational Database Management System) - today's predominant technology for implementing data management environments, which is based upon tabular modeling and storage of data. Studies have discussed data quality assessment for tabular data representation [16, 22]. Parssian et al. [21] offer quantitative methods for assessing the data-defect propagation in RDBMS environments and demonstrate the impact of a poor-quality data-source on the data manufacturing process outcome. A few studies [18, 20, 26] discuss the use of metadata as a possible solution for managing data quality in data warehouses, addressing the specific challenges related to data warehouse environments.

There are several methods for quantitatively assessing data quality described in the literature. Pipino et al. identify 3 archetypes of functional forms: (a) ratio - a proportion between the actually obtained and the expected values, (b) min/max value among aggregations and (c) weighted average between multiple factors [22]. Hufford defines a set of ratio-based quality attributes such as accuracy, completeness and validity [16]. The ratios proposed are based on counts of data entities with quality defects compared to the overall counts. Shankaranarayanan et al. develop ratio and weighed-average definitions for accuracy, completeness and timeliness [24]. Ballou and Pazer base their model on a generalized view of accuracy as a number on a 0-1 scale, where 1 represents error-free data [3]. In a later study, they propose a combination of ratio-based and weighted-average definitions for completeness and consistency [4]. Their study makes a distinction between structural versus content-based quality measurement: the former assumes the existence of an absolute standard and bases the calculation on objective characteristics, such

[1] Analogous to Deming's quality improvement cycle - Plan, Do, Check and Act [10]

as the volume of data recorded. The latter is derived by the information *content*, as it applies to business use by the data consumers. Parssian et al., provide an extensive quantitative analysis regarding the effect of data error propagation through different processing stages on the structural measurement of data quality attributes [21].

Recent studies have further explored the usability of quality measurement as an aid for managerial decision making. Providing users with quality metadata during the decision-making process was empirically shown to impact outcomes [7, 14]. Providing such metadata to business users, together with process metadata (e.g., information about data sources, transformations and storage), can improve their ability to accurately assess data quality, builds a sense of reliability and thus enhance decision-making processes and outcome [25]. Linking quality measurements to business use and managerial decision-making introduces the necessity to distinguish between the impartial and the contextual assessment of data quality. Impartial (objective) refers to the data quality assessment that is derived from the data itself, disconnected from specific usage. Contextual (subjective) refers to quality assessment within the business/decision context accounting for contextual factors. Contextual assessment can be influenced by different factors: (a) *organizational levels:* individuals, departments and the organization as a whole assess data quality differently. An individual business-user, for example, will be more concerned about the data quality of the particular data subsets that he or she uses, while the organizational level IS management, has to look at quality from the broader perspective of the entire data databases collection. (b) *Business tasks:* The characteristics of the task for which the data is used are likely to affect quality assessment. Quality requirements may significantly differ when the data is used for managerial decision making (e.g. wide range of data summarized/aggregated), on-going business operations (e.g. accurate transaction data), or for innovative processes (e.g. accessible external information). (c) *Information stakeholders and stages:* Yang and Strong [33] identify categories of information stakeholders which can be associated with different stages of the DMP and may view and emphasize data quality assessment differently - information collectors who are associated with the data sources (data collection), information custodians who are associated with the technical management of the processes (data storage, transfer and processing), information customers who use the end product, and information managers who are associated with implementation and administration of the entire data manufacturing process. (d) *Timing:* The quality level attributed to data may differ significantly, based on timing of use [3]. For example – the currency of the latest stock price quotes is more critical during business hours when systems are actively used for active trading. (e) *Personal characteristics:* Contextual assessment is also likely to be effected by the personal characteristics of the user – factors like motivation, extent of involvement, and work experience [25]. Wang and Strong show that business users view some quality attributes (e.g. accuracy) as impartial in nature, and others (timeliness, completeness, relevance, believability) as contextual [30]. Both impartial and contextual attributes contribute to the overall perception of quality, hence the importance of assessing quality within context [27]. Shankaranarayanan and Watts propose a dual process approach for information validity assessment, based on quality attributes [25]. According to them, the overall data quality assessment is influenced by both structural and content attributes, mediated by the notion of data relevance and moderated by the user expertise and user involvement.

An observation based on the review of DQM literature is that most of the quantitative methods discussed are structure-based rather than content-based. While structural measurement of quality may be appropriate for impartial assessment, it could be argued that data contents, rather than the structure, influence the business-user's perception of quality. Therefore, contextual assessment will benefit from the development and enhancement of content-based quality measurements methods. This study looks further into this interplay between impartial and contextual quality assessment emphasizing the importance of managing data quality within the context of managerial decision making. To facilitate this, this paper introduces the concept of *intrinsic value* which we define as a conceptual measure of the business value associated with a specific set of data items. In other words, it is a value measure of the importance/worth of a set of data items determined by how much that set contributes to the business activity that it is used for and how much that activity is likely to benefit from that set of data items.

INTRINSIC VALUE

This section introduces the concept of intrinsic value, as a measurement reflecting the relative business value attributed to the data, and describes a quantitative approach for assessing it. The intrinsic value can be used as a baseline for developing a set of value-driven quality measurements, as demonstrated in the following section of this paper. It provides a strong tool for the assessment of data quality in context. The notion of intrinsic value stems from the perception of data as abstracted representation of some business activity. An assumption that underlies the intrinsic value is that the captured data (e.g. a sale transaction or customer information) reflects business activities that are valuable to the organization. While business activities may have similar structural data representation, they are certainly not similar from the standpoint of business-value creation – the more profitable a sale transaction, or the higher lifetime-value of a customer, the greater is the business-value contribution of those transactions (sale or customer records) and associated activities. Data that reflects that activity from a business perspective is more important and has a higher business value. Furthermore, the relative value of the data is likely to be reflected when the data is used for managerial decision making. When the data represents a transaction that is more profitable, the more significant will be the influence of this data on revenue and profitability assessment during the decision-making process. The higher the lifetime-value attributed to a customer, the higher is the likelihood that any data associated with this customer will attract managerial attention for a decision that involves this data. The notion of differentially valuing data in organizations is not entirely new. Customer Relationship Management (CRM) systems help better manage customers by identifying and associating value to sets of customers based on customer transactions and activities (customer data). ABC classification in inventory management is a technique that is based on the fact that not all inventory products are equal; each must be managed differently based on how valuable it is (determined by price, quantity, and turn-over). The associated values (customers or inventory) are not constant, but vary with the changing business environment. Acknowledging the differential valuation of data from the business perspective, this study suggests that the capability of the data to create and reflect business value ought to affect data management, particularly data quality assessment, and promote a broader use of content-based measurement.

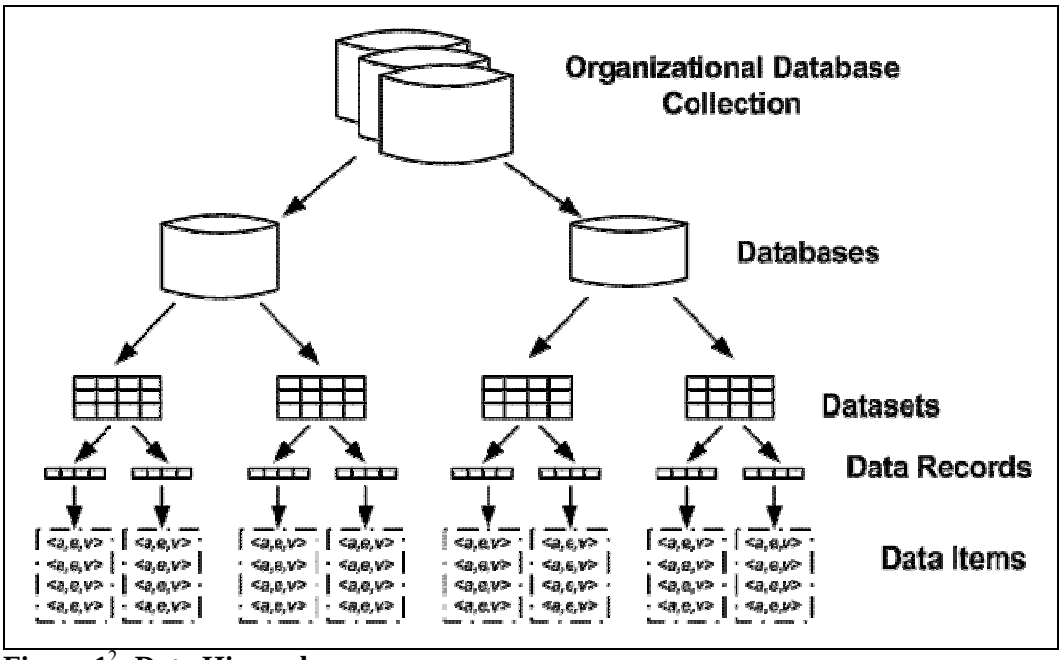

Figure 1[2]: Data Hierarchy

[2] Adapted from [12] and [23]

The intrinsic value and the derived measurements adhere to the hierarchy in data management systems, illustrated in Figure 1. The hierarchy is based upon the *data-item*, or the datum, as the atomic entity. The data-item is defined as a triplet <*a,e,v*> of a value 'v' selected from the value-domain attached to attribute 'a' of entity 'e', which represents a physical or conceptual real-world object. The *data record* is a data-item collection that represents an attribute-set of an object instance. A *dataset* is a record collection that represents the same instance type, and a *database* is a collection of datasets with meaningful relationships among them. Organizations manage multiple databases, *the organizational database collection*, each aimed for different business purposes. This paper focuses on the tabular datasets, which underlie dominant data management technologies (such as RDBMS, delimited text files, legacy systems, spreadsheets, and statistical packages)[3]. The tabular dataset is based upon a two-dimensional table model. Table columns (or fields) represent attributes while rows (or records) represent instances of the modeled entity. Table 1 illustrates a simplified tabular dataset of sales transactions.

ID	***Date***	***Customer Code***	***Product Code***	***Quantity***	***Price***	***Amount***
1	May 2, 2005	C	X	20	$5,000	$100,000
2	May 2, 2005	B	Y	3	$1,000	$3,000
3	May 3, 2005	A	Y	1	$1,000	$1,000
4	May 3, 2005	B	Z	5	$3,000	$15,000

Table 1: Illustrative Sale-Transaction Dataset

The first set of definitions below addresses the lower hierarchical-levels: data-items, data records and datasets. The reason is the repetitiveness within the tabular dataset - having multiple records with identical field structure - allows certain assumptions that are invalid at higher hierarchical levels (database and database-collections. The required adjustments to compute these two are discussed in the following section). The development starts with fundamental definitions for tabular dataset characteristics:

Records ($\{R_n\}_{n=1..N}$): The dataset has N identically-structured records, indexed [n]=1..N

Fields ($\{F_m\}_{m=1..M}$): Each dataset record has M fields, indexed [m]=1..M.

Field Contents ($X_{n,m}$): Field [m] in record [n] has the actual stored value $X_{n,m}$. The value belongs to the value domain attached to the field (e.g., Integer, Real, Alpha-numeric, or a finite set of valid options). Depending on the design, a field may/may not contain a NULL value.

Data Items (Y): The number of atomic data items in the dataset,

(3.1) $$Y = NM$$

Size (S): The size of the dataset in bytes. If the size of field [m] is S_m, the record size S' is:

(3.2) $$S' = \sum_{m=1}^{M} S_m$$

The entire dataset size is given by:

(3.3) $$S = NS' = N\sum_{m=1}^{M} S_m$$

A data record represents a portion of business activities, which may have an embedded notion of value generation to the business. Next, the *intrinsic value* is developed as a relative measurement that reflects the business value that the dataset represents, by defining:

[3] Alternative data modeling and storage approaches (e.g., Object-Oriented, Object-Relational, and XML) are gaining popularity in recent IS implementations. Future studies should look into the possibility of expanding the methods and concepts that are developed here to such alternatives.

Record Value Scaling Factor (K): A fixed, non-negative factor that rescales the intrinsic value to the desired numeric scale. K can be refined by attributing a scaling factor K_m per-field:

(3.4) $$K = \sum_{m=1}^{M} K_m \, .$$

Intrinsic Record Value (V_n): a non-negative real number that represents the relative business value captured by the record. To define this, two high-level computational approaches can be considered:

1. Fixed: Records are assumed to be equally valuable with a fixed intrinsic value V', scaled by K:

(3.5) $$V_n = KV'$$

2. Factored: Business transaction may include non-negative numerical fields (measures) that reflect business activity (e.g., quantity or amount). Such measures can act as a weight-factor for the relative value of the record. Assuming that the numerical field F_k acts as the value weight-factor:

(3.6) $$V_n = KM_n$$

where $M_n=X_{n,k}$, is the value stored in field F_k of record R_n. Factoring can also be based upon alpha-numeric dimension fields that influence value (e.g., a higher lifetime-value attributed to certain customers). M_n is affected by the dimension value, or the combination of dimension values. For example, assigning a weight M=5 to client 'X', M=3 to client 'Y' and M=1 to all other clients.

***Intrinsic Value (V)*:** the dataset-captured business value is the sum of the intrinsic record values:

(3.7) $$V = \sum_{n=1}^{N} V_n \, .$$

More specifically, for the two approaches, fixed and factored, described above:

1. Fixed:

(3.8) $$V = \sum_{n=1}^{N} KV' = NKV' = NV' \sum_{m=1}^{M} K_m$$

2. Factored:

(3.9) $$V = K \sum_{n=1}^{N} M_n = \sum_{n=1}^{N} \sum_{m=1}^{M} K_m M_n$$

Illustrative Example 1: Consider the simplified dataset, shown in Table 1. Assuming a record scaling factor of K = 1, one can examine different methods for calculating the intrinsic value:

Fixed Intrinsic Value: Assigning a fixed value per-record (e.g., V' = 1), the intrinsic value is

(3.10) $$V = NKV' = 4 * 1 * 1 = 4$$

Amount-Factored Intrinsic Value: Using the "Amount" as a scaling factor, the intrinsic value is

(3.11) $$V = K \sum_{n=1}^{N} M_n = 1 * (100{,}000 + 3{,}000 + 1{,}000 + 15{,}000) = 119{,}000$$

Customer-Factored Intrinsic Value: Using the "Customer Code" as a scaling factor, and assigning relative value contribution to each customer (e.g., 1 to A, 2 to B, and 10 to C, based on their customer lifetime value), the intrinsic value is

(3.12) $$V = K \sum_{n=1}^{N} M_n = 1 * (1 + 2 * 2 + 5) = 10$$

The intrinsic value is important and useful as a factor that underlies the calculation of content-based quality measurement, as demonstrated in the following section. The above example shows that the intrinsic value can, but does not necessarily, reflect monetary values. From the standpoint of developing the content-based quality measurements, the intrinsic value represents the *relative* value attributed to a data entity, when compared with other entities within that data collection (e.g., the relative value of a record, or a set of records within a dataset). Hence, going forward the intrinsic value will be viewed as a relative, non-negative assessment of value, with no particular units.

VALUE-DRIVEN QUALITY MEASUREMENT

This section describes data quality measurement methods to better support contextual quality assessment. The attributes demonstrated – completeness, validity, accuracy and currency – are developed for data management environments that are based on tabular datasets. These quality attributes, or dimensions, have been extensively researched in literature. Such measurements belong to the contents quality category (rather then the process, the model, or the presentation), and the development presented here follows the content-based quality measurement, instead of structure-based (as defined in [4]). The quality attributes are first developed for the lower hierarchy-levels (data-item, data record, and dataset), and are later expended to the higher levels (database and database collection). To ensure consistency and usefulness to contextual assessment, the following principles have directed the development of the measurement models:

Interpretation Consistency: Measurements can be defined at each level of the data storage hierarchy – from the granular level of the data item to higher-level data collections such as records, datasets, or databases. Quality measurements, at each level where they are defined, should have consistent semantic interpretation. For example, "Completeness" should have similar semantic interpretation at all hierarchical levels from data-items to entire databases, measuring the extent to which the data collection is not missing contents.

Representation Consistency: The measurement outcome should be easy to interpret by business users. An accepted representation in the DQM literature is to represent data quality measurement as a numeric ratio within the range of 0-1, where 1 represents perfection and 0 represents the poorest possible quality [e.g., 3, 16, and 22]. The 0-1 representation of quality measurements should hold at any hierarchical level, and for any subset of a data collection.

Aggregation Consistency: The calculation of a high-level collection is based on an aggregation over the more-granular components that it contains. The aggregation should result in a 0-1 score that can not be higher than the highest quality level, or lower than the lowest, among the lower-level granular items. Therefore, when all the granular items are of identical quality, the aggregation results in the same score. Aggregation operators that adhere to this concept are, for example, Min, Max, or Weighted-Average [22].

Impartial-Contextual Consistency: The measurement should reflect contextual perception, but has to be derived from impartial characteristics. In a context-free assessment, the calculation should fold-back to the traditional, structure-driven, measurement. For example, with no content-driven aspects present, the content-based "Completeness" should fold back to the ratio between the number of data items recorded, and the number of data items that should have been recorded [as defined in 4].

Completeness (C): A data item is incomplete if it is missing from the dataset or corrupted such that its contents cannot be determined, and complete otherwise. Completeness measures the degree to which the data items in the data collection are complete. Completeness can be defined as a value-scaled ratio at different levels:

(a) Data Item Completeness: $C_{n,m}$, the completeness of the data item stored in field F_m of record R_n, is 1 if the item is complete and 0 if incomplete.

(b) Record Completeness: C_n, the completeness of record [n] is defined as the weighted-average of data item completeness, where the field value scaling factor serves as a weight:

$$C_n = \left(\sum_{m=1}^{M} K_m C_{n,m}\right) \Big/ \left(\sum_{m=1}^{M} K_m\right) = \frac{1}{K}\left(\sum_{m=1}^{M} K_m C_{n,m}\right). \qquad (4.1)$$

(c) Dataset Completeness: C, the dataset completeness, is defined as the weighted-average of record completeness, where the record value scaling factor serves as a weight:

$$C = \left(\sum_{n=1}^{N} V_n C_n\right) \Big/ \left(\sum_{n=1}^{N} V_n\right) = \frac{1}{V}\sum_{n=1}^{N} V_n C_n = \frac{1}{KV}\sum_{n=1}^{N}\sum_{m=1}^{M} K_m V_n C_{n,m}. \qquad (4.2)$$

Specifically addressing the two computation approaches, fixed and factored:

1. Fixed: $V_n = KV'$ and $V = KNV'$, hence completeness is given by

(4.3) $$C = \frac{1}{K^2 NV'} \sum_{n=1}^{N} KV' \sum_{m=1}^{M} K_m C_{n,m} = \frac{1}{KN} \sum_{n=1}^{N} \sum_{m=1}^{M} K_m C_{n,m}$$

2. Factored: $V_n = KM_n$ and $V = K\sum_{n=1}^{N} M_n = \sum_{n=1}^{N} \sum_{m=1}^{M} K_m M_n$, hence

(4.4) $$C = \left(\sum_{n=1}^{N} KM_n \sum_{m=1}^{M} K_m C_{n,m} \right) \Big/ \left(K^2 \sum_{n=1}^{N} M_n \right) = \left(\sum_{n=1}^{N} M_n \sum_{m=1}^{M} K_m C_{n,m} \right) \Big/ \left(K \sum_{n=1}^{N} M_n \right)$$

Since field factors and record factors, when applied, are non-negative – it can be shown based on the definitions above, that completeness is always a number between 0 and 1, where 1 indicates a dataset that is complete and 0 indicates a practically empty dataset. If the factoring assumption is relaxed (equivalent to setting fixed K_m=K and M_n = M), the dataset completeness can be simplified to

(4.5) $$C = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} C_{n,m}$$

Completeness, in this case, is the ratio between (a) the number of complete items: $\left(\sum_{n=1}^{N} \sum_{m=1}^{M} C_{n,m} \right)$, and (b) the total number of items (MN). This ratio between item counts is the traditional, impartial view of completeness defined in literature [4, 16, 23].

Validity (L): A data item is invalid if its contents are not within the pre-specified value domain (including incomplete or a NULL value where not allowed), and is valid otherwise [16]. This has also been termed as integrity or domain consistency [23]. Validity measures the degree to which the dataset has valid items, taking a similar approach as the definition of completeness: The data item validity ($L_{n,m}$) is 1 if the item is valid and 0 if invalid. Expressions for record validity (L_n) and dataset validity (L) can be developed along the same lines as the equations (4.1) through (4.5).

Accuracy (A): A data item is inaccurate if its contents conflict with a baseline that is perceived to be the correct (including incompleteness and/or invalidity), and it is accurate otherwise. Such baseline value, for example, could be the "real world" value of the data item, or a value within another dataset that was reliably validated to be correct. Accuracy measures the degree to which the dataset has accurate items. With data item accuracy ($A_{n,m}$), one can take a similar approach to the definitions of completeness or validity - 1 if the item is accurate and 0 if not. Alternatively, one can consider a more refined approach, using a distance measure:

(4.6) $$A_{n,m} = D\left(X_{n,m}, X^T{}_{n,m}\right)$$

where:

$X_{n,m}$ – The actual value stored in the data item

$X^T{}_{n,m}$ – The correct value that ought to be stored in the data item

D – A distance function, a real number within 0-1. $D(X_{n,m}, X^T{}_{n,m})$ is 1 for $X_{n,m} = X^T{}_{n,m}$, and approaches (or equals) 0 as the error margin between the correct and the actual increases. With numeric fields, for example, the distance can be defined using a declining exponent:

(4.7) $$D\left(X_{n,m}, X^T{}_{n,m}\right) = EXP\left\{ -\alpha \left| X_{m,n} - X^T{}_{m,n} \right| \right\},$$

It can also be defined as a threshold function:

(4.8) $D(X_{n,m}, X^T{}_{n,m}) = 1$, if $|X_{n,m}-X^T{}_{n,m}| < \Delta$, and 0 otherwise.

Once the data item accuracy is defined, expressions for record accuracy (A_n) and dataset accuracy (A) can be developed similarly to (4.1) through (4.5).

Currency (T) measures the degree to which the dataset is current and up to date. In this paper we define it to measure the age of the dataset. Some research studies in DQM have used the attribute *timeliness* as a measure of age [22, 30]. Others use *timeliness* to define on-time availability of the dataset [3, 16]. In this paper we use currency of a dataset as a measure of age and timeliness of a dataset as a measure of on-time availability of that dataset. Currency measurement is based on the data age ($t_{n,m}$), the time-lag between present time and the last update of the data item. Databases can be designed to track the age of each and every data item. However, in reality it is more common to have the age information specified at the record level (e.g. an update date/time field in a tabular database), in which case all data items that are part of the record are assumed to have the record's age, or at the dataset level (e.g. the update date/time of a file), in which case all data items in the dataset are assumed to have the dataset's age. To comply with other quality measurements, data item currency $T_{n,m}$ can be rescaled to a 0-1 range. Possible rescaling formulations, illustrated in Figure 2, are:

1. Exponentially Declining Currency[4]*:* In certain usage scenarios, the quality attributed to a data item declines with age, but an aged data item still has some business value. In such cases, an exponential decline will be an appropriate rescaling formulation:

(4.9) $$T_{n,m}\left(t_{n,m}\right) = EXP\left\{-\alpha t_{n,m}\right\},$$

where

$T_{n,m}$ – The currency of the data item

$t_{n,m}$ - the age of the data item, $t_{n,m} >= 0$

α – an exponential decline factor, $\alpha > 0$. The larger α is, the more rapid is the quality decline with the increase of age.

It must be noted that an exponential rescaling of currency has been proposed by Pipino et al. [22]. They suggest a rescaling factor must be determined by expert analysis.

2. Time-Threshold Bounded Currency: In certain usage scenarios, the quality attributed to a data item declines with age, and at a certain age ($t^*_{n,m}$), the dataset becomes valueless and unsuitable to use. A formulation for such scenarios is:

(4.10) $$T_{n,m}\left(t_{n,m}\right) = \begin{cases} 1 - \left(t_{n.m} / t^*_{n,m}\right)^{1/\alpha} & ,0 \le t_{n,m} \le t^*_{n,m} \\ 0 & ,otherwise \end{cases}$$, where

$T_{n,m}$ – The currency of the data item

$t_{n,m}$ - the age of the data item, $t_{n,m} >= 0$

$t^*_{n,m}$ - the marginal age after which $T_{n,m}=0$

α – a decline factor, $\alpha \ge 0$. $\alpha=1$ indicates a linear decline while $\alpha=0$ indicates a step-function shape: 1 when $t_{n,m} < t^*_{n,m}$, and 0 otherwise.

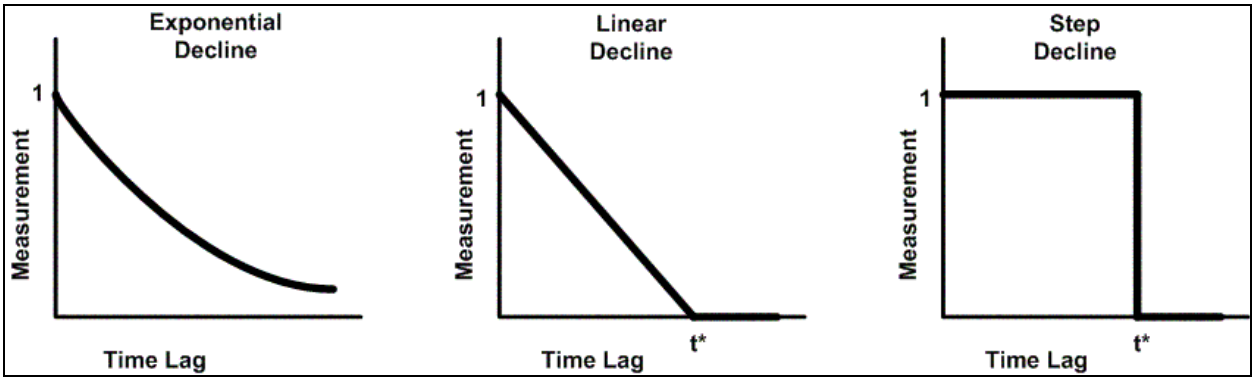

Figure 2: Currency Rescaling Formulations

[6]An exponential rescaling of currency was proposed by Pipino et al. [22], who suggest that the rescaling factor α is to be determined by an expert analysis

Aggregation from the data-item level to the record level (T_n), or to the dataset level (T), can be developed in the same manner as equations (4.1) through (4.5). An alternative aggregation approach is to use the *minimum/maximum* operators [22]: the aggregated age of a collection of items is the maximum of per-item age, or equivalently, the aggregated currency is the minimum of per-time currency.

Illustrative Example 2: Consider the dataset illustrated in table 1. For some reason the first record was corrupted and the data contained in it is unreadable (Table 2 illustrates the dataset that was actually delivered). What is the completeness of the delivered dataset?

ID	*Date*	*Customer Code*	*Product Code*	*Quantity*	*Price*	*Amount*
1	*May 2, 2005*	*C*	*X*	*20*	*$5,000*	*$100,000*
2	May 2, 2005	B	Y	3	$1,000	$3,000
3	May 3, 2005	A	Y	1	$1,000	$1,000
4	May 3, 2005	B	Z	5	$3,000	$15,000

Table 2: The Dataset that was actually delivered

A few different computational approaches for assessing completeness are illustrated as follows:

Structural Completeness: The number of items that should have been delivered is 28 (4 records, 7 fields per records). The actual number of items delivered is 21, hence the completeness in this case is

(4.11) $C = 21/28 = 0.75$

Completeness based upon Fixed Intrinsic Value: Assigning a fixed value per-record (e.g., V'=1), the completeness can be computed as

(4.12) $$C = \frac{1}{KN}\sum_{n=1}^{N}\sum_{m=1}^{M} K_m C_{n,m}$$

In this case, because $C_{n,m}$ is identical for all the fields within a record [n]:

(4.13) $$C = \frac{1}{KN}\sum_{n=1}^{N} KC_n = 0.75.$$

As illustrated, when the intrinsic value per-record is fixed, the result is identical to structural completeness.

Completeness based on Amount-Factored Intrinsic Value: Using the "Amount" field as a scaling factor, the completeness is given by

(4.14) $$C = \left(\sum_{n=1}^{N} M_n \sum_{m=1}^{M} K_m C_{n,m}\right) \Big/ \left(K\sum_{n=1}^{N} M_n\right),$$

or since $C_{n,m}$ is identical within a record [n],

(4.15) $$C = \left(\sum_{n=1}^{N} M_n C_n\right) \Big/ \left(\sum_{n=1}^{N} M_n\right) = 19{,}000/119{,}000 = 0.16.$$

Completeness based upon Customer-Factored Intrinsic Value: Using the "Customer Code" as a scaling factor, and assigning relative value contribution per customer code (e.g., 1 to A, 2 to B, and 5 to C, as in the previous illustrative example), the completeness is given by

(4.16) $$C = \left(\sum_{n=1}^{N} M_n C_n\right) \Big/ \left(\sum_{n=1}^{N} M_n\right) = (1+2*2)/(1+2*2+5) = 5/10 = 0.5$$

The completeness scores are significantly different, depending on the method, but which score is correct? That is determined by and depends on the contextual data use. The structural completeness score best represents the purely-technical perspective – one record out of four is missing, hence a 75% completeness measurement. From accounting perspective, the amount-driven completeness score is probably more relevant – the biggest sale transaction is missing from the database, hence the quality is perceived to be so poor (16%). If the purpose is tracking the activity of key clients, the customer-driven completeness (50%) appears more relevant, since the missing transaction represents a relatively significant customer.

Higher-Level Quality Aggregation: So far, the quality measurements were developed for data collections at lower hierarchical-levels – data items, records, and datasets. A similar aggregation approach can be taken towards developing quality measurements for higher hierarchical-levels. The key for such aggregations is defining a measure of relative value that can be used as a scaling factor. For example one can associate a monetary value to each dataset that can be aggregated to obtain an overall monetary value of the entire database. We define:

Databases ($\{B_q\}_{q=1..Q}$)***:*** Databases managed by the organization, indexed [q]=1..Q.
Datasets ($\{\underline{D}_{q,p}\}_{p=1..P^q}$)*:* Datasets within database [q], indexed [p]=1..P^q
Dataset Value ($W_{q,p}$)***:*** The value of dataset [p] in database [q].
Database Value (W_q)***:*** The total value of datasets in database [q], given by

$$W_q = \sum_{p=1}^{P^q} W_{q,p} \qquad (4.17)$$

Database-Collection Value (W): The total organizational database collection value is given by:

$$W = \sum_{q=1}^{Q} W_q \qquad (4.18)$$

Database Completeness: Assuming that the completeness of datasets in database [q] is given by $\{C_{q,p}\}_{p=1..P^q}$, The completeness of the database, C_q, can be defined by:

$$C_q = \left(\sum_{p=1}^{P^q} W_{p,q} C_{p,q}\right) \Big/ \left(\sum_{p=1}^{P^l} W_{p,q}\right) = \left(\sum_{p=1}^{P^q} W_{p,q} C_{p,q}\right) \Big/ W_q \qquad (4.19)$$

Database-Collection Completeness: The completeness of the entire organizational data collection can be now defined as:

$$C = \left(\sum_{q=1}^{Q} W_q C_q\right) \Big/ \left(\sum_{q=1}^{Q} W_q\right) = \left(\sum_{q=1}^{Q} W_q C_q\right) \Big/ W \qquad (4.20)$$

The other quality attributes – validity, accuracy, and currency – can be modeled at the database level or at the database-collection level using a similar value-based aggregation approach as illustrated above. Such aggregation of quality measurements to the higher hierarchical levels can serve as a useful managerial tool for getting an overall sense of the quality of the organizational data.

IMPLICATIONS AND DIRECTIONS FOR FUTURE RESEARCH

This study introduces a content-based method for data quality measurement, aimed to support contextual data quality assessment. Measurement techniques that use context-independent characteristics are not very useful in decision-making as they do not permit the decision-maker to gauge the quality of the data used in the context of the decision-task in which it is used. Purely contextual characteristics are difficult to quantify. Research has shown that these when combined with context-independent characteristics offer significant benefits in decision environments [14]. The key contribution of this study is the proposed set

of methods for combining context-independent (or impartial, e.g., data contents, field and record count) and context-dependent (the relative intrinsic value) characteristics to *quantitatively* assess data quality measurements. The measurement methods have been demonstrated for four commonly-discussed (in DQM literature) attributes: completeness, validity, accuracy and currency. Transforming the impartial characteristics into context-influenced quality assessment is achieved by weight-averaging along the intrinsic value, and by adding scaling parameters (e.g., $\{K_m\}$, or α in (4.7) and (4.10) respectively). The suggested methods fulfill the consistency guidelines: interpretation consistency was achieved by deriving the attribute at each hierarchical level from the definitions at the level below, starting at the most-granular data-item level. The calculation of all the quality attributes leads to a number between [0-1], hence, representation consistency is achieved. The stated aggregation principles were addressed by using weighted average. Finally, in context-free assessment the measurement folds-back to structural-based calculation hence impartial-contextual consistency is maintained as well. This was demonstrated for the consistency measurement in (4.5), and can be similarly extended to the other attributes.

Content-based quality measurement is argued to better support contextual assessment of data quality which, as pointed out by previous studies, has important implications to business users. Quality information about the provided data, considered a type of metadata, affects the perception of reliability and believability and hence influences the decision outcome. The suggested method has a few degrees of freedom that offer some flexibility when adapting quality metrics to the specific contextual needs - the selection of data attributes from which the intrinsic value is derived, the scaling weights, or the sensitivity factors. As result, the numeric metrics that are provided to end-users are more closely related to content-factors that matter most for business use. Value-based measurement also has important implications from the administration perspective of quality management efforts, as it provides a powerful prioritization tool –data subsets of high perceived relative-value ought to get higher priority and data quality issues in these must be addressed quicker, compared to other datasets.

Value-based computation methods can be integrated into databases, data management systems, and front-end software for supporting data quality management (e.g., [31], [22]). While the calculation methods described here produce numeric measurements that are different from structure-based calculations, they do not have to be presented differently to end-users. The outcome of such calculation, as a number within the range of 0-1, adheres to accepted concepts of quality representation. Such measurements can be perceived as a type of quality metadata and stored as part of the metadata layer, or embedded into data records as additional attributes [26]. Recent studies on the effect of quality metadata on the outcome of decision making call for further investigation regarding the specific set of quality metrics that should be provided to the end-user and the form in which it should be presented [7, 14, 25]. A research direction that is currently under investigation is the synergy-effect between multiple types of metadata and quality metrics. Integration of quality metrics together with metadata information about the data manufacturing process (particularly data sources and data processing methods) may enhance the end-user capability to assess the data quality within a specific context of use, above and beyond the contribution of each component alone.

The value-contribution of data to business, an important aspect that is not very often addressed by DQM research, is highlighted by the value-based measurement. Value contribution of information and IS has long been debated and the Information Value (IV) area, which attempts to address this issue, is among the predominant streams of IS research [6]. IV acknowledges the effect of quality on the value attributed to information. Hilton [15] argues, based on Blackwell's theorem, that among the information value determinants (e.g. set of possible actions, pay-off, and uncertainty) only the IS characteristics, and particularly the information quality, have monotonic relationships with value. However, applying IV principals to data quality management is challenging. First, IV often treats information abstractly, hence the difficulty with adjusting models to the specific characteristics of data management technology. Second, information products have multiple contextual uses hence the overall value is complex to asses.

Third, the complex nature of the DMP increases the difficulty in assessing the value [2]. A possible approach for addressing those challenges is the use of utility functions - a composite of the technology-driven outcome and its translation into business value [15]. Mapping quality measurement, among other IS attributes, into tangible value within a specific usage can provide an important input to value maximization by optimizing design [1]. Utility functions have been discussed in the DQM literature to some extent and their importance of data quality management has been shown [28]. These have been used in DMP optimization methodologies [3, 4, 5]. The intrinsic value and the value-based quality measurements may enhance the benefit gained from the using utility functions for data management. Providing such measurements as input to utility functions has the potential to create a more comprehensible and useful mapping between the data and the business value attributed to its use. Creating such link between the content and quality of the data and the value attributed to it by the business is an important step towards integrating DQM and IV, a synergy that can benefit both fields and introduce a broad range of research opportunities.

REFERENCES

[1] Ahituv N. (1980), *"A Systematic Approach Towards Assessing the Value of Information System"*, MIS Quarterly (4:4), Dec. 1980, pp. 61-75

[2] Ballou D.P., and Pazer H.L. (1985), *"Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems"*, Management Science (31:2), Feb. 1985, pp. 462-484

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

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

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

[6] Bankar R. D., and Kauffman R. J. (2004), *"The Evolution of Research on Information Systems: A Fiftieth-Year Survey of the Literature in Management Science"*, Management Science (50:3), Mar. 2004, pp. 281-298

[7] Chengalur-Smith I., Ballou D. P., and Pazer H. L (1999), *"The Impact of Data Quality Information on Decision-making: An Exploratory Study",* IEEE Transactions on Knowledge and Data Engineering (11:6), 1999, pp. 853-864

[8] Davern M. J., and Kauffman R. J. (2000*),"Discovering Potential and Realizing Value from Information Technology Investments"*, JMIS (16:4), Spring 2000, pp. 121-143

[9] DeLone W., and Mclean E. (1992), *"Information Systems Success: The Quest for the Dependent Variable"*, Information Systems Research (3:1), 1992, pp. 60-95

[10] Deming W.E. (1986), *"Out of Crisis"*, MIT Center for Advanced Engineering Study, Cambridge, MA

[11] Eckerson W.W. (2003), *"Achieving business success through a commitment to high quality data"*, TDWI Report Series, Data Warehousing Institute, Seattle, WA, Http://www.dw-institute.com, 2003

[12] Elmasri R, and Navathe S.B. (1994), *"Fundamentals of Database Systems"*, Second Edition, Benjamin Cummings, Redwood City, CA

[13] English L. P. (1999) *"Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits"*, John Wiley and Sons Inc, New York, NY

[14] Fisher C. W., Chengalur-Smith I., and Ballou D. P. (2003), *"The Impact of Experience and Time on the Use of Data Quality Information in Decision Making*", Information Systems Research (14:2), pp. 170-188

[15] Hilton R.W. (1981), *"The Determinants of Information Value: Synthesizing Some General Results"*, Management Science (27:1), Jan. 1981, pp. 57-64

[16] Hufford D. (1996), *"Data Warehouse Quality"*, DM Review, Jan. 1996

[17] IDC, www.idc.com, Online Reference Journal

[18] Jarke M., Lenzerini M., Vassiliou Y., and Vassiliadis P. (2000), *"Fundamentals of Data Warehouses"*, Springer-Verlag, Heidelberg , Germany

[19] Kahn B. K., Strong D. M., and Wang R. Y. (2002), *"Information Quality Benchmarks: Product and Service Performance"*, Communications of ACM (45:4), Apr. 2002, pp. 184-192

[20] Marco D. (2000), *"Building and Managing the Meta Data Repository: A Full Lifecycle Guide"*, Willey and Sons, Inc., New York, NY

[21] Parssian A, Sarkar S., and Jacob V.S. (2004), *"Assessing Data Quality for Information Products – Impact of Selection, Projection, and Cartesian Product"*, Management Science (50:7), pp. 967-982

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

[23] Redman T.C. (1996), *"Data Quality for the Information Age"*, Artech House, Boston, MA

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

[25] Shankaranarayanan G. and Watts-Sussman S. (2003), *"A Relevant Believable Approach for Data Quality Assessment"*, In the Proceedings of the MIT International Conference on Information Quality (IQ 2003), October 2003, Boston, MA

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

[27] Strong D.M., Yang W.L, and Wang R.Y. (1997), *"Data Quality in Context"*, Communications of the ACM (40:5), May 1997, pp. 103-110

[28] Wand Y., and Wang R. Y. (1996), *"Anchoring Data Quality Dimensions in Ontological Foundations"*, Communications of the ACM (39:11), Nov. 1996, pp. 86-95

[29] Wang R.Y, Storey V. C, and Firth C.P (1995), *"A Framework for Analysis of Data Quality Research"*, IEEE Transactions of Knowledge and Data Engineering (7:4), Aug. 1995, pp. 623-640

[30] Wang R.Y., and Strong D.M. (1996), *"Beyond Accuracy: What Data Quality Means to Data Consumers"*, JMIS (12:4), Spring 1996, pp. 5-34

[31] Wang R.Y. (1998), *"A Product Perspective on Total Quality Management"*, Communications of the ACM (41:2), Feb. 1998, pp. 58-65

[32] Wixom B.H., and Watson H.J. (2001), *"An Empirical Investigation of the Factors Affecting Data Warehousing Success"*, MIS Quarterly (25:1), Mar. 2001, pp.17-41

[33] Yang W.L., and Strong D. M. (2003), *"Knowing-Why about Data Processes and Data Quality"*, Journal of Management Information Systems (20:3), Winter 2003-4, pp. 13-39

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