A Semiotic Information Quality Framework

Rosanne J. Price
Graeme Shanks

School of Business Systems, Monash University
Melbourne, Australia
Email: rosanne.price@infotech.monash.edu.au
Email: graeme.shanks@infotech.monash.edu.au

Abstract

An organization depends on quality information for effective operations and decision-making. However, fundamental questions still remain as to how quality should be defined and the specific criteria that should be used to evaluate information quality. Previous work adopted either an intuitive, empirical, or theoretical approach to address this problem; however, we believe that an integrated research approach is required to ensure both rigour and scope. This paper presents an information quality framework based on semiotic theory, the linguistic theory of sign-based communication, to describe the form-, meaning-, and use-related aspects of information. This provides a sound theoretical basis both for defining quality categories, previously defined in an ad-hoc manner, based on these different information aspects and for integrating the different research approaches required to derive quality criteria for each category. The goal of our work is to provide an approach to defining information quality that is both theoretically grounded and practical that can serve as a basis for further research in data quality assessment and decision support.

Keywords

Information quality, Data quality, Semiotics, Decision support

1. INTRODUCTION

Quality information[1] and information quality management in an organization is essential for effective operations and decision-making. Tactical and strategic decision-making is especially dependent on the quality of the data used in the decision making process. The proliferation of data warehouses as a basis for integrating and aggregating multiple sources of data to support decision making further highlights an organization's vulnerability with respect to poor data quality. The challenge of maintaining a sufficient level of data quality to satisfy business needs is made more difficult in data warehousing and decision support systems in general than in conventional databases supporting business operations. This is due both to the widely disparate data sources, contexts, users, and data uses characterizing data warehouses and the much less predictable data usage involved in decision making as compared to business operations. This raises questions regarding how it is possible to ensure that the quality of data matches the requirements of data uses that are not yet known.

Regardless of whether conventional databases or data warehouses are used to support decision making, it is clear that management of information quality is critical to the effectiveness of the decision support systems employed. However, management of information quality pre-supposes a clear understanding of and consensus with respect to the meaning of the term "information quality". In fact, fundamental questions still remain as to how quality should be defined and the specific criteria that should be used to evaluate information quality. Definitions of *quality* and its associated set of *quality criteria* and *categories* (used to group criteria) found in information system(s) (IS) literature and practice can, in general, be described as coming from either product-based or service-based perspectives and employing either empirical, practitioner, theoretical, or literature-based approaches.

The product-based perspective, commonly called *data quality*, focuses on the design and internal IS view. From this view, quality is defined in terms of the degree to which the IS data meets initial requirements specifications or the degree to which the IS data corresponds to the relevant real-world phenomena that it purports to represent. Typical criteria include *completeness* and *accuracy*, evaluated using objective measures. The

[1] Due to the lack of agreement on the precise definition of *information* in the literature, we choose to restrict our usage of the term *information* to informal discussion (used synonymously with the term *data* unless otherwise noted) and avoid its use in formal definitions. The exception is the deliberate use of the term *information quality* to describe our work because it connotes consideration of data delivery for judging quality.

problem is that even if data corresponds to requirements specifications or the real-world, there can still be quality deficiencies with respect to actual use-related data requirements, which may differ from the planned uses catered for in the initial specifications. In fact, it is evident that information consumers (i.e. internal or external users of organizational data) are the final judge of quality. This leads to a service-based perspective of quality, commonly called *information quality,* which focuses on the information consumer's response to their task-based interactions with the IS. It is this view of quality that directly addresses the question of how to ensure sufficient quality for unpredicted or changed data uses, i.e. by continuous assessment of information consumer quality perceptions. The use of the term *information* rather than *data* implies that the *use* and *delivery* of the data must be considered in any quality judgements, i.e. the quality of delivered data represents its *value* to information consumers using the data. So the IS processes that deliver data to the user are as important as the actual IS data itself in determining quality. From this view, quality is defined as the degree to which the delivered data meets or exceeds information consumer expectations or needs as perceived by the information consumers themselves. Typical criteria include *timeliness*, *relevancy*, and *accessibility* as judged by the information consumers. The single most widely accepted definition of quality is *fitness for use*. Although directly related to the service-based view of quality (i.e. measurement methods and quality judgements may not exactly match those of the product-based view), in some sense it can be seen as subsuming the product-based view as well since any data not meeting product-based requirements is not likely to be judged fit for use.

Information quality research is further characterized by the range of research approaches employed, e.g. empirical, practitioner, theoretical, or literature-based. For example, empirical research approaches such as that by Wang & Strong (1996) rely on information consumer feedback to derive quality criteria and then classify them into categories (defined in an ad hoc manner after criteria derivation). This approach has some important implications: because it is based primarily on information consumer feedback rather than on a systematic theory, there are likely to be some inconsistencies, redundancy, and/or omissions in observable in the list of criteria and their category groupings. In particular, and as previously noted also by Eppler (2001), the quality criteria defined in (Wang & Strong 1996) have significant inter-dependencies (e.g. *believability* subsumes *reputation, ease-of-understanding* subsumes *interpretability*, *conciseness* and *consistency* both contribute to *ease-of-understanding*) that are not explicitly acknowledged or justified. Other listed criteria are not generic, i.e. do not apply across all domains or data types. The criteria *objectivity* illustrates this problem, since some domains require subjective data, e.g. subjective data types such as recorded managerial rankings of goals by priority. Furthermore, the quality categories are not formally defined nor is their selection justified empirically or theoretically. The limited semantic basis for the selection of quality categories and their use in classifying the quality criteria is clear both from (1) the substantial changes evident in category names, explanations, and member criteria in each subsequent paper following on from the original (Huang et al. 1999; Kahn & Strong 1998; Kahn et al. 1997; Kahn et al. 2002; Lee et al. 2002) and (2) from naming ambiguities (e.g. overlapping names such as the *useful* and *effective* categories in (Kahn et al. 1997) or the *access* category and its *accessible* criterion in (Wang & Strong 1996)).

Because the practitioner-based approach is based on ad-hoc observations and experiences, it is similarly subject to criticisms with respect to a lack of rigor. A good example is English's (1999) informal practitioner-based approach to quality, considering both product and service-based perspectives (which he calls *inherent* and *pragmatic*). Although the differentiation between the two categories based on English's static, use-independent versus dynamic, use-dependent quality criteria is quite intuitive; inconsistencies in the classification of quality criteria into the two categories can be clearly observed on the basis of the specified category and criteria definitions. An example of criteria classification by English that contradicts his own stated criteria and category definitions is as follows. Although English defines the quality criterion *precision* as being dependent on data use and the quality category *pragmatic* as being use-independent, he includes *precision* in the *pragmatic* quality category. *Accessibility* is another use-dependent criteria inconsistently classified as *pragmatic*.

There are other problems with English's proposed classification scheme. As with empirically-derived quality definitions, a large number of inter-dependencies are observable between criteria (e.g. most of the *inherent* criteria contribute to the *pragmatic* criterion of *rightness*, *precision* from the *inherent* category contributes to *usability* in the *pragmatic* category) that are not acknowledged or justified. Furthermore, intellectual confusion between desirable quality criteria and the means of evaluating them is evidenced by including *accuracy to a surrogate source* as a criterion separate from *accuracy to reality;* when in fact the former is an attempt to estimate the latter and not a desirable characteristic in its own right.

In contrast, theoretical approaches such as Wand & Wang's (1996) evaluation of correspondence to real-world phenomena derive criteria logically and systematically based on an underlying theory. As a result, the derived quality definitions and criteria generally have a higher degree of rigor and internal coherence as compared to empirical or practitioner approaches. The drawback of this approach is with respect to scope. It is clear that a complete approach to defining quality must take into account suitability for a specific task. Since requirements can never be stated completely and can change over time, any useful judgement of data quality must include some consideration of information consumer perceptions. Since this aspect is not easily amenable to a purely theoretical approach, such approaches are necessarily limited in scope to product-based quality aspects, as acknowledged explicitly by the authors themselves in (Wand & Wang 1996).

Finally, a literature-based approach to quality, based on review and analysis of existing quality literature, is generally not used alone but rather as support for one of the other three approaches of deriving quality criteria or, as in (Eppler 2001), for survey purposes to compare existing quality frameworks.

Thus previous work in defining data quality generally employs either an intuitive practitioner, empirical, or theoretical approach, with support from a literature-based approach. However, we believe that a theoretically-based but integrated approach is required to ensure both rigor and scope. Specifically, we advocate the use of (1) theoretically-based derivation of quality criteria and categories whenever possible for rigor and (2) an underlying theoretical quality framework that concomitantly provides a consistent and thus rigorous basis for integrating other non-theoretical approaches as required for comprehensive scope. This is in accordance with the conclusions from Eppler's (2001) review of information quality frameworks noting the need for a generic (i.e. not domain-specific) and integrated (i.e. combining theoretical and practical aspects) approach.

In this paper, we present a framework for understanding and defining information quality based on semiotic theory, following the approach first proposed by Shanks (1998). As a well-established linguistic theory describing sign-based communication, semiotics can be used to describe the form-, meaning-, and use-related aspects of information and can serve as a theoretical framework to integrate the different approaches required to define quality criteria for each of these different information aspects. The goal of our work is to provide an approach to defining data quality that is both theoretically grounded and practical and that can serve as a basis for further research in data quality assessment and decision support. The specific aim is to provide a theoretical foundation for the later development of practical quality assessment tools and guidelines. Therefore, the focus of this paper is on information quality properties (i.e. quality categories and criteria), with additional consideration given to the quality goals and assessment techniques suitable for each quality category. In Section 2, we review semiotic theory and its application in the IS context and then introduce our approach. This is followed by a description of the semiotic framework in Section 3, including an overview of the framework; definitions of quality categories based on semiotic levels; and a review of quality goals, measurement techniques, and the criteria derivation approach used for each category. Based on this framework, Section 4 has a detailed discussion of the specific quality criteria associated with each individual quality category and discusses their derivation. Finally, Section 5 has conclusions and future work.

2. A SEMIOTIC APPROACH TO QUALITY

The study of signs has been associated with philosophical and linguistic enquiry into language and communication from the time of the Greek philosophers. Modern semiotics, as proposed by Charles Pierce (Pierce 1931-1935)and later developed by Charles Morris (Morris 1938), describes the study of signs in terms of its logical components (*Intl. Enc.Comm.* 1989). These are the sign's actual *representation*; its *referent* or intended meaning (i.e. the implied propositional content, that is, the phenomenon being represented, which may or may not be a physical object), and its *interpretation* or received meaning (i.e. the effect of the representation on an interpreter; its import: in other words, the actual use of the representation in terms of influencing the behaviour or actions of the interpreter). Informally, these three components can be described as the form, meaning, and use of a sign.[2] Relations between these three aspects of a sign were further described by Morris as syntactic (between sign representations), semantic (between a representation and its referent), and pragmatic (between the representation and the interpretation) semiotic levels. Again, informally, these three levels can be said to pertain to the form, meaning, and use of a sign respectively.

The process of interpretation, called semiosis, at the pragmatic level necessarily results from and depends on the use of the sign. This process can be viewed in terms of its potential influence on the interpreter's subsequent actions (i.e. reactions) or, in cases where the sign representation was deliberately generated by a sender, as a means of communication. In either case, the actual interpretation of the sign depends both on the interpreter's general sociolinguistic context (e.g. societal and linguistic norms) and on their individual circumstances (e.g. personal experience or knowledge). With this background, the correspondence between semiotics and information quality can be clarified and the applicability of semiotics to the formal definition of information quality justified.

[2] Note our informal use of the term *meaning* as intended rather than (from Mingers (1995)) as received meaning because it reflects common usage, i.e. is listed as the first definition in all consulted dictionaries.

A datum is maintained in a database or data warehouse precisely because it is representative of some external phenomenon relevant to the organization, i.e. useful for business activities. However, the representational function of the datum is realized only when it is retrieved and used by some entity, where the entity may be either human or machine. Data use necessarily entails a process of interpretation that potentially influences the resulting action taken by the interpreter. For example, a clerk may issue a query and retrieve a stored integer number from a database that they then interpret as the current age of a particular employee. As a result, the clerk then sends a letter to that employee with notification that the employee is approaching mandatory retirement age.

A clear correspondence between the semiotic concept of a *sign* and the IS concept of *datum* can be observed by noting that datum has the same three components described earlier for a sign: a stored *representation*, a represented external phenomenon as the *referent*, and a human or machine *interpretation*. In fact, datum serves as a sign in the IS context. As is true for any sign, the actual interpretation of the representation (and the degree to which that corresponds to the original or intended referent when the sign was generated) will depend on the interpreter's background (i.e. programming for a machine interpreter and societal and personal context for a human interpreter). Similarly, the process of interpreting data can be viewed in terms of its influence on the interpreter's actions or as a form of communication between the sender (e.g. generator of the stored representation) and the user.

There are several precedents for the application of semiotic theory to IS that are relevant to our work. Semiotics has previously been applied to understanding IS by Stamper (1992) in the context of systems analysis, then the related group of papers (Krogstie et al. 1995b; Krogstie et al 1995b; Krogstie 2001; Lindland et al. 1994) in the context of evaluating the quality of data models, and finally by Shanks in the context of evaluating information as well as data model quality (Shanks & Darke 1998; Shanks & Tansley 2002). In particular, Lindland and Krogstie used semiotic levels to classify the different aspects of data model quality and evaluation and Shanks described how a similar approach could be applied to data quality. The work described in this paper follows on from Shanks work to formally define a semiotic information quality framework.

In terms of the semiotic model adopted, our approach differs from earlier applications of semiotics to IS research. As proposed by Stamper (1992) and adopted to varying degrees in (Krogstie et al. 1995b; Krogstie et al. 1995a; Krogstie 2001), an additional three semiotic levels have been used in the IS context. These are social (i.e. shared social context), empiric (i.e. statistical properties of the sign representation), and physical (i.e. physical/material properties of the sign representation). However, there is no theoretical foundation for these additional levels in semiotics. In fact, the concepts described by these additional levels are already covered by the original three levels as follows. With the respect to the social level, the social context is already addressed in the context of semiosis—the process of interpretation—at the pragmatic level in traditional semiotics. The physical and empiric levels have to do with the actual generation and transmission of signs. Although not an explicit focus of Piercean semiotics, sign generation and transmission are implicitly included in the original syntactic level since they describe the process of sign representation in the same way that semiosis describes the process of interpretation at the pragmatic level.

In fact, by treating selected sub-aspects (i.e. pragmatic context, syntactic sign generation and transmission) of the original three semiotic levels as separate levels, the clear distinction between levels is blurred and ambiguity is introduced. For example, the shared social context is part of both the semantic level and the pragmatic level. Furthermore, the original congruence between sign components and semiotic levels is no longer preserved. Therefore, we choose to adhere to the original three semiotic levels defined by Morris (1938).

Given the congruence between the original Piercian semiotics and the concept of information, the syntactic, semantic, and pragmatic semiotic levels can serve as a theoretical foundation for both for (1) defining information quality categories, and for (2) using those definitions to select and rationalize the research approach suitable for deriving each category's quality criteria. This theoretical foundation thus provides a theoretical framework for integrating multiple research approaches and for classifying quality criteria into categories. Because quality criteria are initially derived with reference to a specific quality category based on that category's definition, there is no need for the separate and manual classification of criteria into categories necessary when criteria and categories are derived independently, as is true for other research proposals considering more than one quality category. Instead, the classification is a natural and automatic outcome of the definition of categories and consequent selection of a criteria derivation approach for that category, thus ensuring a consistent classification. This clearly differentiates our work from other information quality approaches. Rather than an ad-hoc and/or empirical derivation of quality categories and classification of quality criteria, the use of semiotics provides a sound theoretical basis for both steps.

In defining a semiotic framework for information quality, our goal was to maintain rigor without sacrificing scope or practicality. As noted in the introduction, theoretical approaches to defining information quality generally have an advantage over other approaches with respect to the level of rigor (e.g. internal consistency, coherency, precision) that can be achieved. The benefit of using semiotic theory with respect to establishing a rigorous basis of deriving quality categories and classifying quality criteria has already been described. However, given the general consensus that information quality cannot be completely assessed without reference to the information consumer (discussed in Section 1); it is clear that not all aspects of information quality are amenable to a purely theoretical approach: information consumer judgments must be considered. Therefore, some degree of empirical work is required to ensure comprehensive coverage of information quality aspects. Thus, an information quality framework should address the need to integrate empirical and theoretical research approaches.

The advantage of using semiotic theory to describe information quality is that the three semiotic levels provide a clear theoretical basis both for (1) identifying and compartmentalizing the quality criteria requiring an empirical approach and for (2) integrating empirical and theoretical approaches within a single coherent information quality framework. Specifically, the pragmatic level requires at least a partly empirical approach since it relates to the use of signs (i.e. data) by interpreters (i.e. information consumers); whereas, a theoretical approach can be used for those levels independent of sign use (i.e. the syntactic and semantic levels). It should be noted that the rigor of empirically-derived quality criteria can be improved with respect to previous empirical work by establishing as a priority from the outset that criteria should be generic and have minimal overlap. The next section describes our semiotic framework in detail.

3. A SEMIOTIC FRAMEWORK FOR INFORMATION QUALITY

In this section, we describe the structure of our proposed information quality framework based on the application of semiotic theory. In adapting semiotics to the IS context, the goal throughout is to adhere as closely as possible to the original structure and definitions of semiotics as espoused by Charles Pierce (1931-1935) and Charles Morris (1938), thus ensuring the consistency and legitimacy of the adaptation. Most importantly, as discussed in the Section 2, we adopt the original three semiotic levels and their definitions as described by Morris. Each of these levels is then used to formally define a separate category of information quality criteria with its associated quality goals, evaluation technique, and criteria derivation approach.

Before discussing the semiotic levels and derived information quality categories, we first present the semiotic definition of a sign and its application in the IS context to the definition of data and meta-data. Note that although the term *real-world* is commonly used to describe what is represented by data; we prefer to use the terms *external* to refer to represented manifestations. This term is more general in that it includes phenomena outside the real-world (e.g. the supernatural or imaginary mental constructions such as unicorns) and allows us to make the distinction between data and meta-data. In addition, references to stored data assume a single abstract IS representation of the different levels of actual physical representation (i.e. essentially nested signs).

Definition 1. A *sign* is a physical manifestation (i.e. the representation) with implied propositional content (i.e. the referent) that has an effect on some agent (i.e. the interpretation, resulting in some behaviour, either action or understanding, by the agent).

Definition 2. A *datum* is a sign in the IS context with a stored form (i.e. the representation); an intended meaning which is the represented external phenomenon relevant to a specific application domain (i.e. the referent); and a human or machine interpretation involving some use of the datum (i.e. the interpretation).

Definition 3. A *meta-datum* is a sign in the IS context with a stored form (i.e. the representation); an intended meaning (i.e. the referent), that is, a represented rule that constrains or a document that describes either external phenomena relevant to a specific application domain or data organization relevant to a specific data model; and a human or (usually) machine interpretation involving some use of the meta-datum by the human or machine agent (i.e. the interpretation).

Definition 4. *Data* is a set of datum collected based on their shared relevance to achieve some goal, which may involve a set of tasks directed towards a common broad goal.

Definition 5. *Meta-data* is a set of meta-datum in the IS context specifically collected for the purpose of organizing, documenting, and/or constraining data in an IS.

Essentially, data and meta-data comprise the contents of a database or data warehouse. They both serve as signs in the IS context representing respectively external phenomena relevant to an application or external rules or documentation relevant to an application or data model. For example, meta-data include business integrity rules constraining the combinations of data values that are legally allowed in the database or data warehouse (i.e. based on application rules describing possible external states) and general integrity rules constraining the data organization in the IS (i.e. based on the underlying data model employed by the IS). For the sake of clarity, we will refer to these two classes of integrity rules as *application* and *data model* integrity rules respectively. To illustrate the above discussion, the application integrity rule *employee.age<65* serves as a sign for the business rule that existing employees must retire when they are 65 and only employees under that age are hired. Similarly, the data model integrity rule *employee.dept# is a foreign key for dept.dept#* serves as a sign for the specific relational referential integrity rule that a specified employee's department must exist. In other words, meta-data include the set of definitions (and documentation) relating to either the business application domain or to the underlying data model that form the IS design. Next, the definitions of each semiotic level and the derived information quality categories are given.

Definition 6. The *syntactic level* consists of any relation between sign representations.

Definition 7. The *semantic level* consists of any relation between a sign representation and its referent.

Definition 8. The *pragmatic level* consists of any relation between a sign representation and its interpretation.

Definition 9. The *syntactic quality category* describes the degree to which stored data conform to stored meta-data.

Definition 10. The *semantic quality category* describes the degree to which stored data corresponds to represented external phenomena, i.e. the set of external phenomena relevant to the purposes for which the data is stored (i.e. use of the data).

Definition 11. The *pragmatic quality category* describes the degree to which stored data is suitable and worthwhile for a given use, where the given use is specified by describing two components: an activity (i.e. a task or set of tasks) and its context (i.e. location—either regional or national—and organizational sub-unit—typically created as a result of functional, product, and/or administrative sub-division).

Definitions 6-8 for the semiotic levels and definitions 9-11 for the resulting derived quality categories relate respectively to the form, meaning, and use of signs. Essentially the syntactic and semantic categories relate to the product-based and the pragmatic category to the service-based quality views described in Section 1. In practice, definitions 9-11 would be applied with respect to a specific data set. The syntactic and semantic quality categories have a direct correspondence to the definition of their respective semiotic levels. For example, since data and meta-data are both signs in the IS context; the conformance of stored data (e.g. employee John's stored age of 55) to stored meta-data (e.g. the stored rule that employee age must be less than 65) describes a relation between sign representations. Similarly, the correspondence of stored data to represented external phenomena describes relations between sign representations and their referents. In defining the *pragmatic quality category*, we have focused on one aspect of the interpretation as described in Section 2, i.e. the use of the representation. Thus the relation between stored data and its use describes relations between sign representations and the aspect of interpretation related to their use. In the context of information quality, *use* is further described by both activity and context, since any judgement regarding the suitability and worth of a data set are dependent on both aspects of use.

To summarize, the three semiotic levels—*syntactic, semantic,* and *pragmatic*—describing respectively (1) form, (2) meaning, and (3) application (i.e. use or interpretation) of a sign can be used to define corresponding quality categories based respectively on (1) conformance to database rules, (2) correspondence to external (e.g. real-world) phenomena, and (3) suitability for use. Using the example of an employee database, these three quality aspects can illustrated by (1) no employee records having an age attribute of more than 65, assuming that such an integrity rule has been defined, (2) a given employee record (or set of records based on foreign-key-based relational joins) correctly represents an real employee (e.g. has matching details), and (3) employee information available in the database is useful for the tasks performed by the information consumer.

Table 1 below shows the ideal and operational quality goals, quality evaluation technique, and quality criteria derivation approach for each quality category. The table further shows the quality question addressed (for reference purposes) and the level of objectivity for each quality category.

The ideal and operational quality goals for each quality category follow directly from the definitions of that category, where the operational goals differ from the ideal goals in that they allow a user-specified degree of deviation from the ideal. With respect to the syntactic and semantic categories, this allowable deviation entails specification of an acceptable error rate (i.e. data not conforming to integrity rules or corresponding to external phenomena). With respect to the pragmatic category, the allowable deviation entails specification of an acceptable gap between expected versus perceived quality. This is based on service quality theory, described in (Parasuraman et al. 1988; Parasuraman et al. 1991) with subsequently proposed variants described in (Dyke et al. 1997; Pitt et al. 1997). Service quality theory employs a difference score or gap, evaluated by surveying customers, to measure their quality perceptions of services rendered (e.g. car repair, travel booking).

The next row of the table describes the evaluation techniques relevant to assessing each quality category. From this description, it can be seen that the derivation of quality criteria for each quality category could be used to support development of an automated integrity-checking tool at the syntactic level, sampling guidelines at the semantic level, and a questionnaire used to solicit information consumer feedback at the pragmatic level.

The effectiveness of data quality evaluation techniques at the syntactic level depends on the quality of the metadata (i.e. integrity rules); however, since that is outside the scope of this paper we assume perfect metadata.[3] Integrity checking then entails using automated techniques to check data for conformance to the integrity rules specified, for example, as data declarations, triggers, or active rules.

Sampling techniques used to evaluate quality at the semantic level require that both external phenomena and data be sampled to assess the degree of incomplete (missing) and spurious (un-matched) representation respectively. If it is impractical to access the external phenomena directly, a trusted surrogate such as a telephone directory for people's names and addresses may be employed instead as an approximation.

	Syntactic	**Semantic**	**Pragmatic**
Quality Question Addressed	Is IS data good relative to IS design (as represented by metadata)?	Is IS data good relative to represented external phenomena?	Is IS data good relative to actual data use, as perceived by users?
Ideal Quality Goal	complete conformance of data to specified set of integrity rules	1:1 mapping between data and corresponding external phenomena	data judged suitable and worthwhile for given data use by information consumers
Operational Quality Goal	user-specified acceptable % conformance of data to specified set of integrity rules	user-specified acceptable % agreement between data and corresponding external phenomena	user-specified acceptable level of gap between expected and perceived data quality for a given data use
Quality Evaluation Technique	integrity checking, possibly involving sampling for large data sets	sampling using selective matching of data to actual external phenomena or trusted surrogate	survey instrument based on service quality theory (i.e. compare expected and perceived quality levels)
Degree of Objectivity	completely objective, independent of user or use	objective except for user determination of relevancy and correspondence	completely subjective, dependent on user and use
Quality Criteria Derivation Approach	theoretical, based on integrity conformance	theoretical, based on a modification of Wand and Wang's (1996) ontological approach	empirical, based on initial analysis of literature to be refined and validated by empirical research

Table 1: Quality Category Information

In terms of the relative objectivity of the different quality categories, it can be seen from Table 1 that the degree of objectivity decreases from syntactic to semantic to pragmatic categories. A comparison of stored data to stored metadata at the syntactic level is completely objective, since it depends only on what data and integrity rules are currently stored. The semantic category involves some degree of subjectivity, since comparing stored data to relevant external phenomena requires that relevancy and correspondence judgements be made. In general, one would expect a large degree of consensus in these judgements. However, these determinations do involve some subjectivity, since the purposes for which the data is stored and the way external phenomena are represented may by understood differently by different individuals or change over time. This is especially true in the case of data warehouses and other decision support systems used primarily for tactical and strategic rather than operational business process support, since it is difficult to predict how the data will be used or to preserve the original data context and both are very likely to evolve over time. Finally, the pragmatic category involves completely subjective judgements based on user perceptions of quality.

Table 1 further indicates the approach used to derive quality criteria for each category. Theoretical techniques can be used for both syntactic and semantic categories; however, empirical techniques are required for the pragmatic level because it depends on information consumer quality judgements relating to the data's suitability and worth for a given data use. That is, although an initial set of pragmatic quality criteria can be

[3] In fact, metadata will always be imperfect, due to limitations in expressivity of existing integrity languages and because it is not practical to completely specify the applicable set of integrity rules.

proposed based on an analytic review of quality literature, they require validation through empirical research methods. In the next section, we discuss the derivation of quality criteria for each category in detail.

4. DERIVING QUALITY CRITERIA FOR EACH QUALITY CATEGORY

Regardless of the approach used to derive quality criteria, there are several requirements and goals that can be formulated prior to and considered throughout the derivation process to ensure a systematic and rigorous evaluation of potential quality criteria. The requirements are as follows:

- quality criteria must be general, i.e. applicable across application domains and data types,
- quality criteria must clearly defined,
- quality criteria must be expressed as adjectives to ensure consistency, and
- overlap (i.e. interdependencies) between criteria must be fully documented and justified.

The goals are as follows:

- the names of quality criteria should be intuitive, i.e. corresponding as closely as possible to common usage,
- quality criteria should be non-overlapping (i.e. not have dependencies on other criteria), and
- the set of quality criteria should be comprehensive.

These are listed as goals rather than requirements since there may be circumstances where the first two goals may not be completely satisfiable and we cannot prove that the last goal is satisfied—it can only be subjectively assessed over time through peer review and empirical feedback.

4.1 Syntactic Quality Criteria

The single syntactic quality criterion of *conforming to metadata* is derived directly from the definition of the syntactic quality category, i.e. the conformance of data to metadata.

1. *Conforming to Metadata:*
   Data obeys the constraints described by the specified integrity rules.

This quality criterion can be illustrated by the example given earlier in Section 3 for the syntactic quality category, i.e. no employee records have an age value more than 65, assuming such an integrity rule was specified.

4.2 Semantic Quality Criteria

The derivation of semantic quality criteria are based on the work by Wand and Wang (1996), because it is unique in the quality literature for its theoretical and rigorous approach to the definition of quality criteria. As acknowledged by the authors, the scope of their paper is limited to what they term *intrinsic* criteria—data quality criteria based on the stored data's fidelity to the represented real-world rather than based directly on data use. However, their definition of *intrinsic* quality criteria corresponds to that of our *semantic* quality category. Therefore, their work can serve as a basis for the derivation of semantic quality criteria.

Essentially, Wand and Wang's approach entails a systematic examination of possible design and operational mapping deficiencies that can arise during the transformation from real-world states to IS representations of those states and derivation of quality criteria based on that analysis. This assumes an ontological view that the IS represents the real-world application domain, and that both are composites described by states and laws governing allowed states and their transitions. The transformation process from real-world to IS (including both IS design and the data generation component of IS operations) and from IS back to the real-world (the data retrieval part of IS operations) is described as *representation* and *interpretation* transformations respectively. Based on an analysis of possible data deficiencies arising during the representation transformation phase, Wand and Wang conclude that there are four intrinsic quality criteria that must be satisfied to ensure that the IS is a proper representation of the real-world. That is, the mapping must be:

- complete, i.e. every legal real-world state (e.g. in the application domain) can be represented in the IS
- unambiguous, i.e. no two legal real-world states map into the same IS state (or equivalently, a given IS state infers at most one real-world state),

- meaningful, i.e. every IS state infers at least one legal. real-world state (or equivalently, no meaningless IS states exist that do not infer any legal real-world state), and
- correct, i.e. the representation of a real-world state by an IS state (i.e. mapping from a real-world to IS state) is such that the inference (i.e. reverse mapping) from IS to real-world recovers the original real-world state rather than a different (i.e. legal but not identical to the original) real-world state.

As stated by the authors, their work was intended to be used as a guide for system design and data production to ensure quality IS design and operation; whereas our work is intended to serve as a basis for information quality assessment. To address both the difference in goals and two observed flaws in the original analysis described below, the original analysis and resulting list of derived quality criteria are amended as follows.

- Representation versus Interpretation Phase Analysis of Deficiencies.
  Wand and Wang's analysis of possible mapping deficiencies is based on the representation transformation phase, thus supporting their goal of guiding system design and data production. Since the representation transformation phase includes not only operational (i.e. data generation) but also design aspects (i.e. IS design), this means that the design-based aspects of the data deficiency analysis relate specifically to data model rather than to information quality[4]. In contrast, to support our goal of facilitating information quality assessment, it is more useful to analyse possible deficiencies from the interpretation phase since information quality assessment involves judgements regarding existing data. This ensures that the primary focus is on information rather than data model quality. Furthermore, such an approach simplifies the analysis since there is no need to consider all of the possible causes of an identified discrepancy. Instead, one need only identify all the types of discrepancies that could possibly be observed when retrieving data. These are just the cases of missing, incorrect, ambiguous, meaningless, or redundant data noted by Wand and Wang, amended as discussed in the points below.
- Tautological Definition of *Incomplete* and Concomitant Omission of a Possible Data Deficiency.
  Wand and Wang defines *incomplete* as the case where a real-world state *cannot* be represented by the IS and ascribes this to an error in design. However, this deduction is tautological since the definition's wording limits its applicability to design failures. As a result, the other possible cause of incomplete data was omitted, i.e. missing data as the result of an operational rather than design failure. An example would be when a data entry clerk manually entering data into the IS accidentally omits an entry. A more correct and general definition of *incomplete* (especially from the view of the interpretation transformation perspective) is where a real-world state *is not* represented by an IS, either because it cannot be (i.e. design error) or can be but is not (i.e. operational error). Note that this implies that information quality assessments based on existing data (and its use) might therefore not detect design flaws that potentially could lead to a data deficiency (such as incomplete data) until real-world states that are affected by the design flaw materialize. This can be understood in terms of the differential focus on data model versus information quality discussed in point 1 above.
- Inconsistent Treatment of Meaningless and Redundant States.
  While Wand and Wang recognize that an IS with *meaningless* states can still represent the real-world adequately, they consider it a case of poor design and thus classify it as a data deficiency. On the other hand, they do not classify *redundancy* as a data deficiency even though they note that it could potentially lead to a deficiency, and further claim that it is not at all related to design decisions. We would argue to the contrary that redundancy is traditionally considered a design issue (e.g. whether to enforce the uniqueness constraint of primary keys or whether to duplicate data at distributed sites for efficiency). In fact, we view the case of meaningless and redundant states as quite similar in nature. That is, although neither necessarily results in a deficient real-world representation; they each have a significant potential to lead to data deficiencies (i.e. if meaningless data is accidentally interpreted as a real-world representation or if redundant data is updated inconsistently). Therefore, we feel that these two cases should be treated consistently. Specifically, we conclude that both *meaningful* and *non-redundant* should be considered data quality criteria, while acknowledging that they differ from other semantic criteria in that they represent a danger rather than a definite deficiency.

Further points of difference that should be noted but do not substantively affect the analysis of possible mapping deficiencies or derivation of criteria are as follows.

- IS and Real-World States versus IS Artefacts and Real-World Phenomena.

[4] Information quality commonly refers to the quality of IS data rather than the data model, two distinctly different concepts requiring separate analysis and treatment. The focus throughout this paper is on the quality of data.

Wang and Wand's definitions are expressed in terms of the states (i.e. the state of the entire database compared to the state of that portion of the real-world to be represented); however, that is not practical for information quality assessment. We express our definitions in terms of database and external (e.g. real-world) states, but operationalize the definitions by substituting IS data and real-world phenomena (whose states can be sampled individually). As discussed by Wand and Wang, these two perspectives are interchangeable when analysing data deficiencies, except in the special case of *decomposition deficiencies*. In this case, the overall IS state may not correspond to the real-world even though individual components do, as a result of differently timed update of individual components. In practice, this means that sampling of individual IS and real-world components may not entirely suffice to estimate correspondence: some degree of aggregation may be required to detect decomposition deficiencies.

- Reference to Real-World versus External Phenomena.
  Wand and Wang's paper is framed in terms of comparisons with the real-world, which implies that imaginary and supernatural phenomena are excluded; whereas, they are included in ontologies such as that described by Chisolm (1996). Therefore, we use instead the more inclusive term *external* to support the broadest range of IS data types.

- Assumption of Perfect Analysis and Implementation.
  Wand and Wand analyse the possible source of data deficiencies only in terms of design and operational sources of failure, based on the assumption of perfect analysis and implementation. Because we focus on the interpretation rather than the representation phase, we are concerned only with operational evidence of data discrepancies based on existing data and do not make any assumptions regarding possible failure sources (i.e. causes). In fact, data discrepancies could arise from analysis, design, implementation, or operational (including maintenance) failures.

- Claim that Intrinsic Quality Criteria are Use-Independent.
  Wand and Wand claim that their derived criteria are completely use-independent. However, as discussed in Section 3 with respect to semantic quality, we would argue that the assessment of such criteria is partly dependent on use insofar as the selection of the set of external states to be used for comparison to database states is use-dependent.

We now present the list of semantic quality criteria, based on the *intrinsic* quality criteria defined by Wand and Wang with the amendments as described above:

1. *Complete:*
   Each external state maps to at least one IS state.

2. *Unambiguous*:
   Each IS state maps to no more than one external state.

3. *Correct*:
   The mapping of external to IS state is such that the reverse mapping preserves the original details of the external state

4. *Non-redundant*:
   Each external state maps to no more than one IS state.

5. *Meaningful:*
   Each IS state maps to at least one external state.

Note that, in the current context, we prefer the term *correct* to the commonly used terms *accurate* or *consistent* because of the latter terms' inappropriate connotations relating to numerical precision and uniformity respectively. Notice further that the emphasis is on existing rather than legal external and IS states since the definitions are intended to serve as the basis for information quality assessment and thus from the view of the interpretation rather than the representation transformation. As discussed earlier to further support quality assessment, these state-based definitions can be operationalized in terms of individual external and IS phenomena (e.g. an individual real-world object represented by a separately identifiable database unit such as a single relational record, a set of relational records derived from a join operation, or an object in an object-oriented database) which can then be sampled. However, as noted earlier, some aggregate sampling may be required to detect decomposition deficiencies.

Together, the first 3 quality criteria express the minimal semantic quality requirement that each external state map to at least 1 IS state and each IS state map to no more than 1 external state. The full set of 5 criteria further restrict the mapping to be exactly 1-to-1, i.e. each external state maps to exactly 1 IS state and that each

IS state maps to exactly 1 external state. This represents the optimal semantic quality requirement. To illustrate these 5 semantic quality criteria, we use the employee database introduced earlier and operationalize the definitions as described above. An employee database is *complete* if all the actual employees are represented, *unambiguous* if each employee record can be mapped to only one actual employee, *correct* if the details (i.e. field values) of each employee record in the database match the corresponding properties of the represented employee (e.g. the *sex* field value matches that of the actual employee represented by the employee record), *non-redundant* if each actual employee is represented only once in the database, and *meaningful* if every employee record matches to at least one actual employee.

4.3 Pragmatic Quality Criteria

Based on the definition of the pragmatic quality category, all pragmatic quality criteria should relate to data use, i.e. are evaluated with respect to a specific activity and its context. That implies that the assessment of such criteria will be based on information consumer perceptions and judgements, since only they can assess the quality of the data relative to use. Thus, although an initial list of pragmatic quality criteria can be constructed on the basis of an analysis of current quality literature, validation and refinement of this list is ultimately dependent on empirical feedback from information consumers. The goal of such a validation and refinement process is to identify and correct any omissions, extraneous inclusions, ambiguity, or previously unidentified inter-dependencies in the list of pragmatic criteria. In this section, we discuss the initial list of pragmatic criteria and their rationale. Empirical validation is in progress and will be reported in future publications.

Before considering existing quality literature to construct an initial list of quality criteria, we note that the information consumer's perceptions of the quality criteria listed earlier at the syntactic and semantic levels are also of importance in a comprehensive judgement of quality. For example, the sampling-based assessment of *completeness* at the semantic level may result in an objectively high score; whereas, the information consumer may perceive the level of completeness as unacceptably low in relation to the stringent requirements of their particular use of the data. Thus the information consumers' subjective and use-based judgements may differ from objective and relatively use-independent measurements of the same quality criteria. This represents additional quality information which must be included to fully understand the quality of an organization's data. Therefore, in order to assess information consumer perceptions of syntactic and semantic criteria, these criteria should be included as separate criteria at the pragmatic level.

The approach taken in selecting the initial list of pragmatic quality criteria was to review existing quality literature and analyse proposed criteria to determine first which are use-related—regardless of the individual author's original categorization—and then to examine them for any overlap, inconsistencies, or omissions, in-order to determine a minimally redundant yet comprehensive list. It was observed that many of the specific criteria listed could be seen as specific aspects of more general quality criteria, e.g. that *timeliness, format*, *degree of precision* all related to *suitability*. Thus the large number of criteria related to use could be reduced to a manageable number and, in some cases, inter-dependencies eliminated, simply by explicitly grouping such criteria. As noted by Eppler (2001), this is important because too large a number of either quality categories or criteria within those categories makes the classification more difficult to remember and hence less practical to use. The initial list of pragmatic criteria resulting from this process is as follows:

1. *Conforming to Rules:*
   Data obeys business and other integrity rules.
2. *Reliable:*
   Data corresponds to (i.e. is a trustworthy representation of) relevant external phenomena.
   Sub-dimensions: correct, unambiguous, meaningful, non-redundant.
3. *Complete:*
   The collection of data (i.e. data extent) includes all the information needed for your use of this data.
4. *Understandable:*
   Data is presented in a manner easy to interpret.
5. *Accessible:*
   Data is easy and quick to retrieve.
   Sub-dimensions: easy to access, quick to access.
6. *Secure:*
   Data is appropriately protected from damage or abuse (including unauthorized access).
7. *Flexibly Presented:*
   Data can be easily manipulated and the data presentation customized as needed.

Sub-dimensions: easy to aggregate, easy to change (i.e. convert) units, precision, or representation.

8. *Suitably Presented:*
   The data is presented in a manner appropriate for your use of this data (i.e. your work).
   Sub-dimensions: timely, suitably formatted, suitably precise, suitably measured (with respect to units).

9. *Relevant:*
   The types of data available (i.e. data intent) are pertinent to your use of this data.

10. *Valuable:*
    The data is useful and sufficient for (i.e. important for) your use of this data.

The first 3 criteria relate to information consumer *perceptions* of the syntactic and semantic category criteria. In order to facilitate information consumer understanding of criteria that they evaluate, the original syntactic criterion *conforming to metadata* and its definition was re-worded in terms of *rules,* i.e. to refer only to database integrity rules rather than metadata. For the same reason, the more general criterion *reliable* was used in place of the original more specific semantic criteria *correct, unambiguous, meaningful,* and *non-redundant*—now listed as sub-dimensions. The specific criteria are somewhat technical, involving an understanding of mapping constraints that may be difficult for some information consumers to comprehend. Therefore, the term *reliable* was used instead as it is more intuitively understandable and can be used to represent (i.e. to group) the specific criteria. The quality criterion *reliable* subsumes the commonly listed and inter-dependent quality criteria *believable*, *reputable*, and *accurate* found in the literature. The original semantic criterion *complete* was further amended for use in the pragmatic quality category as follows.

The original definition of the semantic criterion *complete* was directly relevant to the correspondence between data and external phenomena described by the semantic quality category. In the context of the pragmatic category's use-based quality criteria, *complete* refers to whether the extent of data available is sufficient for the use of the data. The definition of the pragmatic criterion *complete* was thus changed accordingly. Note that, in the quality literature, sometimes the term *complete* is used also to refer to whether the types of data available are sufficient for the use of the data. However, this overlaps with the ninth criterion *relevant,* commonly included with *complete* in the list of quality criteria, which relates to whether the types of data available are appropriate for the use of the data. Essentially, this definition of *complete* refers to an aspect of relevance, since if some data types required for a particular use of the data are missing then the data is less relevant for that use of the data. Therefore, we employ the more restricted definition of the quality criterion *complete* since the aspect of completeness relating to types is subsumed by the definition of the quality criterion *relevant*. This has the twofold advantage of eliminating an inter-dependency between quality criteria commonly observed in the quality literature and preserving a definition of *complete* close to that used for the semantic category. However, in order to avoid any confusion or ambiguity in interpretation, it is important that this distinction between the definitions of *complete* and *relevant* be made clear to the information consumer evaluating the criteria.

The next 5 criteria (i.e. 4-8) relate to data presentation and delivery, i.e. data as a service. Two of these criteria, *flexibly presented* and *suitably presented*, represent groupings of related criteria commonly found in quality literature.

Finally, the 10th criteria *valuable* relates to the overall worth or importance of the data with respect to the use of that data. Of all the quality criteria listed, this is the most problematic in that it has inter-dependencies with all of the other quality criteria. That is, data which is not highly rated with respect to other criteria (e.g. not complete, not reliable) will necessarily be less valuable as a result. However, in accord with most information quality researchers, we believe that a comprehensive understanding of quality requires the inclusion of such a criterion, sometimes termed *value-added* or *value*. In particular, even data rated highly in terms of all the other listed quality criteria may still be deficient with respect to quality aspects specific to a given application domain or organizational context. In essence, the quality criterion *valuable* acts as a generic place-holder for those aspects of quality specific to a given application rather than universally applicable. Thus, other than replacing the generic quality criterion with the appropriate domain-specific terms for each individual application, the only other option is its inclusion despite the resulting inter-dependencies. The problems and significance of this particular quality criterion has not, to our knowledge, previously been acknowledged in the literature. However, we believe, as previously noted by Eppler (2001), that explicit recognition of inter-dependencies between quality criteria is an important pre-requisite for quality assessment since the inter-dependencies may have implications for the analytic methods used in the evaluation.

The first 9 pragmatic criteria can be illustrated using the employee database described earlier, assuming the perspective of an administrative employee responsible for generating and sending employee pay checks. If the retrieved salary per pay period for an employee exceeds the specified maximum, then this would not be seen as *conforming to rules*. Complaints to the employee regarding incorrectly addressed or missing pay checks may result from problems with the *reliability* (e.g. incorrect employee address retrieved) or *completeness* (e.g. missing employee record) of stored information and thus affect the employee's perceptions regarding these quality criteria. Employee addresses using non-standard abbreviations may compromise *understandability.* Password protection for sensitive salary information may contribute to overall *security* by preventing unauthorized access. However, from the perspective of the employee regularly responsible for accessing the salary information, the accompanying user verification process may be seen to degrade *accessibility* by increasing access time. Thus we can see that the quality criteria may involve trade-offs. The salary information for an employee is *suitably presented* for pay check generation if it is given per pay period. If not initially so specified, it may still be *flexibly presented* if it is relatively easy to select such an option (e.g. via an interface) or easily manipulate the data to calculate the salary per pay period (e.g. via a pre-programmed function). We can see that the information consumer's quality judgements will be affected not only by the actual stored data but also by the interface used to access that data. Employee salary and address information is *relevant* to generating and sending pay checks respectively; whereas, employee birth date is not.

Finally, to illustrate the concept of domain-specific *valuable* data, we use an example from a specialized domain: spatial data relating to a survey of regional land parcel boundaries. For such data to be assessed with respect to quality or to be regarded as high quality, the lineage of the data—involving details regarding data capture and transformations—must be known (i.e. associated with the spatial data in question). Thus the value of the data depends on its lineage, i.e. the general quality criterion *valuable* acts as a placeholder for the specific spatial quality criterion of lineage.

5. CONCLUSION

In this paper, we have defined an information quality framework based on concepts from semiotic theory to provide a rigorous theoretical foundation for (1) deriving and defining quality categories; (2) classifying and deriving quality criteria; and (3) integrating different research approaches to deriving quality criteria. In particular, three different quality categories relating to data form, meaning, and use were defined based on the syntactic, semantic, and pragmatic semiotic levels respectively. These definitions were then used to justify the selection of a particular research approach as suitable to derive quality criteria for a given category. Quality criteria for the syntactic and semantic levels were derived using a theoretical approach, employing respectively integrity theory and Wand and Wang's (1996) ontologically-based analysis of real-world to data mapping deficiencies. The initial list of pragmatic quality criteria was derived based on an analysis of current quality literature and empirical validation and refinement is in progress. With one syntactic, five semantic, and ten pragmatic quality criteria, the quality categories and the final list of sixteen quality criteria proposed for each is shown in Table 2 below, where the first three pragmatic dimensions are user-based perceptions of syntactic and semantic criteria.

	Syntactic	**Semantic**	**Pragmatic**
Quality Category Definition	data conformance to metadata (i.e. database rules)	data correspondence to external phenomena	data worth (importance) for use
Quality Criteria	conforming to metadata (i.e. database rules)	complete, unambiguous, correct, non-redundant, meaningful	perceived rule conformance, perceived reliability, perceived completeness based on data use, understandable, accessible, secure, flexibly presented, suitably presented, relevant, valuable

Table 2: Proposed Quality Criteria by Quality Category

The described benefits consequent on the adoption of semiotic theory as the theoretical foundation of the proposed framework represent contributions that clearly differentiate our work from other work in information quality. In particular, common problems observed in existing information quality frameworks with respect to inconsistencies in category definitions and criteria classification are addressed by providing a theoretical basis for these steps that naturally supports consistency. Similarly, the establishment and presentation of requirements and goals guiding the quality criteria derivation process ameliorate identified sources of problems or limitations in existing information quality frameworks as discussed in Section 1. These include application-specific, inter-dependent, and inconsistently or ambiguously defined or named quality criteria.

The work described in this paper represents the first phase of a continuing project intended to develop theoretically-based information quality assessment techniques and tools. The first phase was devoted to the definition of a comprehensive information quality framework that could serve as a basis for such development, requiring further only the initial validation and refinement of the pragmatic criteria based on focus group research. This task is currently in progress. The next phase of the project involves the development of an assessment instrument for pragmatic quality and its validation by empirical field test, to be followed by an examination of sampling and integrity checking techniques and tools for assessing semantic and syntactic quality respectively. In summary, by providing a theoretical foundation for the definition of information quality, the work reported in this paper can serve as a basis for theoretically-grounded information quality assessment or decision support research.

REFERENCES

Dyke, T. P. V., Kappelman, L. A. & Prybutok, V. R. (1997), 'Measuring Information Systems Service Quality: Concerns on the Use of the SERVQUAL Questionnarie', *Management Information Systems Quarterly,* vol. 21, no. 2, pp. 195-208

Eppler, M. J. (2001), 'The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks', *Studies in Communication Sciences,* vol. 1, pp. 167-182

Huang, K.-T., Lee, Y. W. & Wang, R. Y. (1999), *Quality Information and Knowledge*, Prentice Hall PTR, Upper Saddle River, New Jersey.

International Encyclopedia of Communications (1989), ed. Erik Barnouw, Oxford University Press, Oxford.

Kahn, B. K. & Strong, D. M. (1998), 'Product and Service Performance Model for information Quality: an Update', in *Conference on Information Quality*, Cambridge, MA, USA, pp. 102-115.

Kahn, B. K., Strong, D. M. & Wang, R. Y. (1997), 'A Model for Delivering Quality Information as Product and Service', in *Conference on Information Quality*, Cambridge, MA, pp. 80-94.

Kahn, B. K., Strong, D. M. & Wang, R. Y. (2002), 'Information Quality Benchmarks: Product and Service Performance', *Communications of the ACM,* vol. 45, no. 4, pp. 184-192

Krogstie, J. (2001), 'A Semiotic Approach to Quality in Requirements Specifications', in *IFIP 8.1 Working Conference on Organizational Semiotics*, eds. Stamper et al., Montreal, Canada, pp. 231-249.

Krogstie, J., Lindland, O. I. & Sindre, G. (1995a), 'Defining Quality Aspects for Conceptual Models', in *IFIP8.1 working conference on Information Systems Concepts (ISCO3): Towards a Consolidation of Views*, eds. Falkenberg, E. D., Hesse, W. & Olive, A., Marburg, Germany, pp. 216-231.

Krogstie, J., Lindland, O. I. & Sindre, G. (1995b), 'Towards a Deeper Understanding of Quality in Requirements Engineering', in *7th International Conference on Advanced Information Systems Engineering (CAiSE'95)*, eds. Ivari, J., Lyytinen, K. & Rossi, M., Springer Verlag, Jyväskylä, Finland, pp. 82-95.

Lee, Y. W., Strong, D. M., Kahn, B. K. & Wang, R. Y. (2002), 'AIMQ: a Methodology for Information Quality Assessment', *Information & Management,* vol. 40, pp. 133-146

Lindland, O. I., Sindre, G. & Sølvberg, A. (1994), 'Understanding Quality in Conceptual Modeling', *IEEE Software*, pp. 42-49

Morris, C. (1938), 'Foundations of the Theory of Signs', in *International Encyclopedia of Unified Science*, vol. 1, University of Chicago Press, London.

Parasuraman, A., Berry, L. L. & Zeithaml, V. A. (1991), 'Understanding Customer Expectations of Service', *Sloan Management Review,* vol. 32, no. 3, pp. 39-48

Parasuraman, A., Zeithami, V. A. & Berry, L. L. (1988), 'SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality', *Journal of Retailing,* vol. 64, no. 1, pp. 12-40

Pierce, C. S. (1931-1935), *Collected Papers*, Harvard University Press, Cambridge.

Pitt, L. F., Watson, R. T. & Kavan, C. B. (1997), 'Measuring Information Systems Service Quality: Concerns for a Complete Canvas', *Management Information Systems Quarterly*, pp. 209-221

Shanks, G. & Darke, P. (1998), 'Understanding Data Quality in Data Warehousing: A Semiotic Approach', in *MIT Conference on Information Quality*, eds. Chengilar-Smith & Pipino, L., Boston, pp. 247-264.

Shanks, G. & Tansley, E. (2002), 'Data Quality Tagging and Decision Outcomes: An Experimental Study', in *IFIP Working Group 8.3 Conference on Decision Making and Decision Support in the Internet Age*, Cork, pp. 399-410

Wand, Y. & Wang, R. Y. (1996), 'Anchoring Data Quality Dimensions in Ontological Foundations', *Communications of the ACM,* vol. 39, no. 11, pp. 86-95

Wang, R. Y. & Strong, D. M. (1996), 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems,* vol. 12, no. 4, pp. 5-34

COPYRIGHT