24
A Semiotic Information Quality Framework: Development and Comparative Analysis

Rosanne Price and Graeme Shanks
Faculty of Information Technology, Monash University, Australia

Introduction

Quality information and information quality management in an organization is essential for effective operations and decision-making. The proliferation of data warehouses to support decision-making further highlights an organization's vulnerability with respect to poor data quality, especially given 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.

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. Addressing these research questions is an important step in establishing a basis both for developing information quality assessment mechanisms and for discussing related issues such as quality improvement and management.

Competing views of quality from *product*- and *service-based* perspectives focus on *objective* and *subjective* views of quality, respectively.

Reprinted from "A semiotic information quality framework: development and comparative analysis," by R. Price and G. Shanks in *Journal of Information Technology*, 20, 2005, pp. 88–120. With kind permission from the Association for Information Technology Trust. All rights reserved.

L. P. Willcocks et al. (eds.), *Enacting Research Methods in Information Systems*

Objective measures of information quality can be based on evaluating data's conformance to initial requirements specifications and specified integrity rules or its correspondence to external (e.g. real-world) phenomena. However, such a view of quality overlooks aspects, critical to an organization's success, related to data delivery and presentation, actual data use, and information consumer perceptions, where an *information consumer* is defined as an internal or external *user* of organizational data.

Actual operational use of data may differ substantially from that considered during system development as a result of omitted, unanticipated, or changing business requirements. This may, for example, result in deficiencies in data model quality (a separate topic on its own, but not the focus of this paper) with respect to actual user requirements, leading to consumer perceptions of poor information quality. Furthermore, even if data meet basic requirements; data judged to be of good quality by objective means may be regarded as inferior by consumers either due to problems resulting from data delivery (e.g. deficient delivery mechanisms, processes, or interfaces) or due to customer expectations that exceed basic requirements.

To address these concerns, subjective measures of information quality can be used based on consumer feedback, acknowledging that consumers do not (and cannot) judge the quality of data in isolation but rather in combination with the delivery and use of that data. Thus, data delivery and use-based factors are integral to a service-based view and to consumer perceptions of quality. The obvious challenge of this approach compared to the objective approach is the difficulty in reliably measuring and quantifying such perceptions.

Note that objective versus subjective views of quality reflect commonly discussed IS distinctions between the terms *data* and *information,* distinguishing between what is stored (i.e. stored data values) and what is retrieved from data collections (i.e. received data values). In this paper, the term *data* is used specifically to refer to stored database or data warehouse content; whereas the term *information* is used in a broader sense to include not only stored data but also *'received'* data that have been delivered to, presented to, and interpreted by the user. Thus the term *information quality* refers to both objective views of stored data quality and subjective views of received data quality. Information quality research can then be characterized based on the view(s) of quality considered.

Information quality research is further characterized by the range of research approaches employed, that is, empirical, intuitive (i.e. *ad hoc),* theoretical, and/or literature-based. Although some authors, for example, Eppler (2001), have used the term *theoretical* to describe approaches based on review and analysis of existing quality literature; we distinguish explicitly between theory-based and literature-based approaches. The intuitive or *ad hoc* approach can be based on industrial experience, common sense, and/or intuitive understanding. A number of frameworks have been proposed in recent years for information quality (Redman, 1996; Wand and Wang, 1996; Wang and Strong, 1996; Kahn *et al.,* 1997, 2002; English, 1999; Lee *et al.,* 2002) based on these different approaches. A detailed comparison of these frameworks (and the research approach adopted by each) with the one proposed in this paper is given in the penultimate section. Here we highlight the steps involved in developing such a framework, the research approaches used, and their limitations – thus providing a motivation for the research reported in this paper.

An information quality framework typically consists of a set of quality criteria and their definitions grouped into general categories that have been separately defined. Even in the case of restricted frameworks that consider criteria from only one category (Wand and Wang, 1996), the different quality categories are initially delineated and defined before restricting the scope. In general, the steps, implicit or explicit, required in developing an information quality framework can be described as follows:

- derivation and definition of quality categories,
- selection of the derivation approach to use for deriving criteria,
- derivation and definition of quality criteria, and
- classification of criteria into categories.

Most notably, what all of the frameworks proposed to date have in common is a non-theoretical basis for these steps. Because the development of the frameworks thus depends (either directly or indirectly) on information consumer feedback or *ad hoc* observations and experiences rather than on a systematic theory, the resulting frameworks are likely to have some inconsistencies, redundancy, and/or omissions and thus are subject to criticism regarding the degree of rigor. This is particularly true of the definitions of quality categories and the subsequent classification of criteria based on these categories. Such inconsistencies have in fact been noted previously by Eppler (2001) and Gendron and Shanks (2003) among others. In general, the only exception to these observations is the theoretical and thus rigorous basis provided by

Wand and Wang (1996) for selecting a derivation approach for objective criteria and for deriving and defining those criteria. However, not only is their scope limited to the objective view of information quality (and thus does not consider subjective quality criteria) but also their initial delineation of categories is intuitive rather than theoretical.

These observations motivate the search for a different approach to developing an information quality framework – one that maintains rigor, especially with respect to the definition of quality categories and classification of criteria into categories, without sacrificing scope, that is, which incorporates both product and service quality perspectives in one coherent framework. This paper reports on an information quality framework, *InfoQual,* developed with these goals in mind. Previously published papers (Price and Shanks, 2004, 2005) have reported in detail on specific aspects of or developmental stages in the research. Here, we present the essential elements of the developmental process as a whole – both theoretical and empirical phases – in the context of a detailed comparison to other information quality frameworks with respect to the developmental approach adopted and consequent implications for consistency and scope.

The development of the framework can be described in terms of five steps:

1. defining quality categories, covering both objective product and subjective service quality views,
2. determining the derivation method to use for criteria in each category based directly on the definition of that category, which effectively provides an automatic and natural classification of criteria into categories,
3. deriving the criteria for the objective product quality component(s) of the framework,
4. deriving the criteria for the subjective service quality component(s) of the framework, and
5. empirically refining the criteria, especially subjective criteria, using focus groups. Note that this step does not involve any re-classification of criteria, since a sound basis for criteria classification is established based on category definitions as described in step 2.

To ensure rigor, a theoretical approach was used wherever possible, that is, in the first three steps. The first two steps were based on semiotics; whereas the third step employs database integrity theory and mapping cardinalities (based on an ontological view of an IS). This raises the question of scope. To be comprehensive, an information quality framework must include subjective component(s) that depend on information consumer judgments both with respect to establishing the relevant set of quality criteria to consider and with respect to assessing quality based on these criteria. Such components are obviously not amenable to a purely theoretical approach. Therefore, the set of subjective service quality criteria were initially derived using a literature-based approach and then – to ensure relevance – empirically refined and validated.

The rest of the paper is structured as follows: The following section reviews semiotic theory and its application in an Information Systems (IS) context. The next section then describes how semiotic theory is used to derive and define quality categories and to determine (and thus justify) the research approach employed for deriving quality criteria. The initial derivation of specific criteria for each category is explained in the subsequent section and their refinement, based on empirical feedback from focus groups, is reported in the section thereafter. The revised framework is presented in the penultimate section with a detailed comparison to previously proposed frameworks. The final section describes conclusions and future work.

Semiotics

Although semiotics has many different branches, the one most relevant in the current context is that proposed by Charles Pierce (1931–1935) and later developed by Charles Morris (1938). In particular, Morris describes the study of signs in terms of its logical components (Barnouw, 1989). These are the sign's actual *representation;* its *referent* or intended meaning (i.e. the phenomenon being represented); and its *interpretation* or received meaning (i.e. the effect of the representation on an interpreter's actions, that is, the actual use of the representation). Informally, these three components can be described as the form, meaning, and use of a sign. 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 by the interpreter. 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.

A datum is maintained in a database or data warehouse precisely because it is representative of some external[1] (e.g. real-world) phenomenon relevant to the organization, that is, useful for business activities. However, the representational function of the datum is realized only when it is retrieved and used by some entity, 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 a 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, a 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 referent originally intended in sign generation) will depend on the interpreter's background (i.e. programming for a machine interpreter and societal and personal context for a human interpreter).

Precedents for the application of semiotic theory to IS include the application of semiotics to understanding IS and systems analysis (Stamper, 1991), to evaluating data model quality (Krogstie *et al.,* 1995; Krogstie, 2001), and to evaluating information quality (Shanks and Darke, 1998). Following Stamper's lead, these authors introduce additional semiotic levels not supported by semiotic theory that (1) introduce overlaps obscuring the clear distinction between levels (e.g. both the pragmatic level and the newly introduced social level address shared social context) and (2) do not preserve the original congruence between sign components and semiotic levels described above. Therefore, we choose instead to adhere to the original three semiotic levels defined by Morris.

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 (1) defining information quality categories, (2) using those definitions to select and rationalize the research approach suitable for deriving each category's quality criteria, and (3) categorizing quality criteria. In fact, it is important to note that the last step follows implicitly (i.e. automatically) from the first two, ensuring consistent criteria classification. Since 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. 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.

A semiotic view of quality categories

In this section, we describe the basic structure of the information quality framework InfoQual in terms of quality categories derived from the three semiotic levels. The intention throughout is to give an informal description sufficient to serve as a basis for understanding the rationale for and structure of the framework. A detailed description of the theoretical development with formal definitions of all the terms is found in Price and Shanks (2004).

We begin by presenting the relevant IS terminology used and its equivalents in semiotic terms. Essentially, *data* and *metadata* together 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 and external rules or documentation relevant to an application or data model. For example, metadata 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, e.g. employee age must be less than 65 years) and general integrity rules constraining the data organization in the IS (i.e. based on the underlying *data model* employed by the IS, e.g. the referential integrity rule that an employee department must exist). In other words, metadata include the set of definitions (and documentation) relating to either the business application domain or to the underlying data model that together form the IS design.

Having established the congruence between IS and semiotic constructs, the definition of information quality categories based on semiotic levels follows naturally. The *syntactic* and *semantic quality categories* have a direct correspondence to the definition of their respective semiotic levels. For example, since data and metadata are both signs in the IS context; the conformance of stored data (e.g. employee John's stored age of 55 years) to stored metadata (e.g. the stored rule that employee age must be less than 65 years) describes a relation between sign representations. Similarly, the correspondence of stored data (e.g. John's stored age) to represented external phenomena (i.e. John's actual age) describes relations between sign representations and their referents. In defining the *pragmatic quality category,* we focus on one aspect of the interpretation as described in the previous section, that is, 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 in terms of a specific activity, its context, and user characteristics; since any judgement regarding the suitability and worth of a data set are dependent on these aspects of use. Note further that references to *stored data* assume a single abstract IS representation of the hierarchically structured logical and physical representations (e.g. files, records, fields, bytes, bits) of IS internals, which can be considered an example of nested signs. Given these explanations, the quality categories can then be defined with respect to a given data set as follows.

Definition 1. The *syntactic quality category* describes the degree to which stored data conform to stored metadata. This category addresses the issue of quality of IS data relative to IS design (as represented by metadata) and is assessed through integrity checking.

Definition 2. The *semantic quality category* describes the degree to which stored data correspond to (i.e. map to) represented external phenomena, that is, the set of external phenomena relevant to the purposes for which data are stored (i.e. the intended use of the data). This category addresses the issue of the quality of IS data relative to represented external phenomena and is assessed through random sampling.

Definition 3. The *pragmatic quality category* describes the degree to which stored data are suitable and worthwhile for a given use, where the given use is specified by describing three components: an activity, its context (i.e. geographic or organizational), and the information consumer characteristics (i.e. experience, knowledge, and organizational role). This category addresses the issue of the quality of IS data relative to actual data use, as perceived by users, and is assessed through the use of a questionnaire or survey.

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

Table 24.1 Application of semiotics to IS: semiotic theory and IS equivalent

Theory: semiotic level	**Application: IS equivalent**
Syntactic level (sign form) sign ← → sign	Data in database conform to integrity rules? For example, *emp.salary* > 0 or *emp.dept#* = *dept.dept#*
Semantic level (sign meaning) sign ← → referent	Data in database match external phenomena? For example, *emp* attribute values match real-world employee details
Pragmatic level (sign use) sign ← → use	Data in database useful for tasks? For example, include details needed for payroll

quality categories based respectively on (1) conformance to database rules, (2) correspondence to external phenomena, and (3) suitability for use. This is illustrated in Table 24.1 using the example of an employee database.

Essentially, the syntactic and semantic categories relate to the objective product-based and the pragmatic category to the subjective service-based quality views described in the first section. The advantages of having a single framework incorporating both views of quality is that it (1) provides a comprehensive description of quality and (2) facilitates comparison between different quality perspectives. In the context of quality assessment, such comparisons can be used to check for discrepancies between objective and subjective assessment methods that are likely to signify a quality problem and may facilitate analysis into the source of the quality problem.

Next, we consider the derivation of quality criteria for each category. As stated earlier, the goal is a general understanding of the approach adopted for each category.

Deriving quality criteria for each category

Regardless of the approach used to derive quality criteria, there are several requirements and goals that were 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:

- criteria must be general, that is, applicable across application domains and data types, and
- criteria must be expressed as adjectives (or adjectival phrases) to ensure consistency.

The goals are as follows:

- the names of quality criteria should be intuitive, that is, corresponding as closely as possible to common usage,
- criteria must clearly defined,
- inter-dependencies between criteria should be minimized as far as possible and, where unavoidable, should be fully documented and justified, and
- the set of criteria should be comprehensive.

These are listed as goals rather than requirements since we cannot prove that these goals are satisfied – they can only be subjectively assessed over time through peer review and empirical feedback.

Theoretical techniques are used to derive quality criteria for both syntactic and semantic categories, as described in the following two consecutive subsections, respectively. The initial list of pragmatic criteria is derived based on an analysis of current information quality literature, as described in the 'Pragmatic criteria' section. The summarized list of initial criteria for each category is given in the subsequent subsection.

Syntactic criterion

The syntactic criterion of *conforming to metadata* (i.e. *data integrity rules)* is derived directly from the definition of the syntactic quality category based on integrity theory. Note that although in the most general theoretical sense metadata comprises definitions, documentation, and rules (i.e. the data schema); we operationalize the definition in terms of conformance to specified integrity rules to serve as a practical basis for syntactic quality assessment. In the context of relational databases, this would comprise general integrity rules relating to the relational data model (e.g. domain, entity, and referential integrity) and those integrity rules specific to a given business or application.

Semantic criteria

The derivation of semantic quality criteria is based on the work of 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 the objective view of quality based on the stored data's fidelity to the represented external world (i.e. not on data use). However, this corresponds to our definition of the semantic quality category; so their work can serve as a basis for deriving semantic quality criteria.

The derivation of quality criteria in Wand and Wang is based on an analysis of possible data deficiencies arising during the transformation of real-world states to IS representations, assuming an ontological view that the IS represents the real-world application domain. Using the example of an employee database, a *good* representation of the real-world by an IS requires that the IS data be *complete* (i.e. not missing anything, e.g. all employees are represented), *unambiguous* (i.e. maps uniquely to the real-world, e.g. a given stored employee ID does not map to two different employees), *meaningful* (i.e. no spurious or unmapped data, i.e. no extra invalid stored employee IDs), and *correct* (i.e. corresponds, e.g. stored employee ID and details match that of the actual employee to be represented). These criteria and their definitions were amended as described in Price and Shanks (2004) to account for differences in goals and to remedy observed inconsistencies in the original analysis. Here we discuss only the two amendments that are directly relevant to the discussion of focus group results and resulting revision of framework criteria (including semantic criteria) in the section on 'Practitioner, academic, and end-user focus groups' and the penultimate section, respectively.

Wand and Wang's original definitions are expressed in terms of database and real-world states; however, that is not practical for information quality assessment. Instead, the definitions must be operationalized in terms of identifiable *IS data units* (consisting of one or more data items, e.g. relational records with fields) and *external phenomena* (e.g. represented real-world objects) whose states can be sampled individually. As discussed in 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, in addition to sampling individual IS and real-world components, some degree of aggregation may be required to detect decomposition deficiencies.

Finally, Wand and Wang classify only *meaningless* but not *redundant* IS states as a mapping deficiency. This is despite the acknowledgement that either case has a significant potential to lead to data deficiencies. We felt that these two cases should be treated consistently. Specifically, we concluded that *meaningful* and *nonredundant* should both be considered information quality criteria, while acknowledging that they differ from other semantic criteria in that they represent a danger rather than a definite deficiency. This issue is revisited in the 'Criteria definition' section as a result of focus group feedback.

Pragmatic criteria

Having reviewed the derivation of quality criteria for the syntactic and semantic quality categories, we next consider the derivation of quality criteria for the pragmatic quality category. Theoretical derivation techniques were suitable for the first two quality categories. However, as described in the introductory section, a combination of literature-based (described in this section) and empirical (described in the next section) techniques are required for the pragmatic category because it relates to consumer use of data and thus is subject to information consumer judgement. The initial set of pragmatic level criteria, described next, were thus derived based on an analytic review of literature guided by the goals and requirements for quality criteria described in the beginning of this section.

Pragmatic criteria pertain either to the delivery or to the importance of the retrieved data. They address the ease of retrieving information *(accessibility),* the degree to which the presentation of retrieved information is appropriate for its use *(presentation suitability),* the comprehensibility of presented information *(understandability),* the ease of modifying the presentation to suit different purposes *(presentation flexibility),* the degree of information protection *(security),* the importance and sufficiency of information for consumer's tasks *(value),* and the relevance of information to consumers' tasks *(relevance).* The last criterion, *relevance,* relates specifically to the types of information available (i.e. *data intent)* rather than to the quantity of information available (i.e. *data extent),* since the latter is already covered by the semantic quality criterion *complete. Value* (i.e. the criterion *valuable)was* included despite acknowledged inter-dependencies with other criteria, because it was considered necessary to act as a generic placeholder for those aspects of quality specific to a given application domain. This is discussed further in the section on 'Inter-dependencies between criteria'.

The pragmatic category includes additional criteria addressing consumer *perceptions* of the syntactic and semantic criteria described earlier. These are included because an information consumer's subjective and use-based judgement may differ considerably from objective and relatively use-independent measurement of the same quality criterion.

Table 24.2 Quality criteria by category

Syntactic criteria (based on rule conformance)	Conforming to metadata, i.e. data integrity rules
Semantic criteria (based on external correspondence)	Complete, unambiguous, correct, non-redundant, meaningful
Pragmatic criteria (use-based consumer perspective)	Accessible (easy, quick), suitably presented (timely; suitably formatted, precise, and measured in units), flexibly presented (easily aggregated; easily converted in terms of format, precision, and unit measurement), understandable, secure, relevant, valuable Perceptions of syntactic and semantic criteria

An example is that the *completeness* of a given data set may be rated as quite good based on an objective, sampling-based semantic-level assessment but may be considered unacceptably poor by those consumers whose particular use of the data impose unusually stringent requirements.

Summarized list of criteria

Table 24.2 presents the initial list of quality criteria derived for each level as described in this section, with any sub-criteria listed in parentheses.

In the next section, we describe the empirical research method used to refine the framework, particularly with respect to the pragmatic criteria.

Practitioner, academic, and end-user focus groups

The primary motivation for conducting focus groups was to refine the initial list of pragmatic criteria derived through an analytic and literature-based approach. The necessity of using such a combined approach was explained in the Introduction, that is, empirical techniques are required to solicit consumer input as to the appropriate set of pragmatic quality criteria since by definition they relate to the subjective consumer perspective. The choice of empirical technique adopted was based on the highly interactive nature of focus groups (Krueger, 1994), allowing for a full exploration of relevant (and possibly contentious) issues based on a direct exchange of views between participants. Such consumer input implicitly provides some indirect evaluation of syntactic and semantic criteria and the framework as a whole, since some of the pragmatic criteria are based on perceptions of syntactic and semantic criteria.

Three focus groups were conducted to solicit feedback from IT practitioners, IT academics, and end-users respectively. The practitioner focus group had eight participants including both data management/quality consultants and in-house IT professionals at varying levels of seniority (i.e. from application developers to senior managers). The academic focus group consisted of seven academics whose research was in the area of data management. The end-user focus group of six participants included administrative, managerial, and technical database users with non-IT backgrounds. Participants were asked to complete an individual opinion form evaluating the pragmatic criteria prior to their attendance at a focus group discussion of those criteria and of related quality issues.

During the focus group discussion, participants were passionate about their views and experiences of quality issues and the challenges of ensuring quality. The wideranging discussion that ensued addressed topics such as defining, assessing, improving, and managing quality in organizations. Since the framework was presented as intended to serve as a basis for development of quality assessment techniques and tools, the relevance of the framework to quality assessment was a major focus of the discussion – especially for practitioners and academics. In this paper, we report those focus group results (based on both individual opinion forms and the group discussion) most directly impacting the InfoQual framework revision. Relevant focus group outcomes can be categorized as related either to missing criteria or sub-criteria, interdependencies between criteria, criteria definition, or framework context and are discussed in the following four subsections, respectively.

Missing criteria or sub-criteria

Participants suggested a number of potential additions to the list of quality criteria. Some were determined to be outside the scope of the framework. For example, the proposed criteria *data model quality* (i.e. metadata quality) is a distinct topic requiring separate analysis and treatment, as discussed in the Introduction. Although poor metadata quality can negatively impact information quality (i.e. be a source of poor information quality), the two terms are not synonymous. Other proposed additions were already covered by the original set of quality criteria. For example, privacy issues related to unauthorized access, use, or distribution of data can be addressed through a minor amendment to the definition of the existing quality criteria *secure,* as discussed in the section ‘Criteria definition’.

Only one proposed addition was both within the framework scope and not currently addressed, *allowing access to relevant metadata*. As it is clearly use-related, this criterion is added to the pragmatic category. In fact, the suggestion to include this quality criterion arose more than once in feedback from separate focus groups and was motivated by the requirements of different application contexts, including the following:

- for documentation on version and update lag time of replicate data,
- for currency, lineage, granularity, transformation, and source documentation of spatial data,
- for documentation on data collection purposes to comply with privacy legislation, and,
- for documentation of context for data originating from disparate or unfamiliar sources, for example, as in a data warehouse or data collection external to the organization accessing the data.

A data set that does not provide access to relevant metadata may result in the data being unintelligible, misinterpreted, or unintentionally misused. This clearly impacts the perceived quality of the retrieved information. It further implies inter-dependencies that should be acknowledged between this new quality criterion and the criteria *understandable* and *secure*.

In terms of sub-criteria, concerns were raised by endusers regarding the level of detail considered by the *correct* criterion. They regarded it as extremely important to differentiate between the specific types of errors that could result in a violation of this quality criterion, that is, a mismatch between a database value (i.e. attribute field value) and the external property that it was supposed to represent. For example, recording that the errors were due to missing values could allow explicit evaluation of the percentage of missing values for a given type of attribute (i.e. field). Possible types of errors include a missing value (i.e. empty field), an inappropriate value (i.e. of the wrong type or an invalid type, e.g. an address or numeric value in the employee name field), or an appropriate but incorrect value (i.e. of the correct type but not matching the external property, e.g. an address value in the address field but not that of the relevant employee). These three types of errors can be re-phrased in positive terms as sub-criteria of the *correct* criterion as: *present* (i.e. field has a value), *appropriate* (i.e. field value is of the correct type), and *matching* (i.e. field value matches that of the external property represented) respectively. Note that these sub-criteria are not independent since *matching* implies *appropriate* which in turn implies *present.* Note also that the presence of NULLs in the database could potentially obscure such judgements unless their meaning is clearly defined. Of the four possible interpretations of NULL described by Redman (1996), *not applicable* and *none* would not violate the *correct* criterion, *applicable but unknown* would violate the sub-criterion *present,* and *applicability unknown* cannot be evaluated (i.e. might be an example of any of the other three cases). These sub-criteria are discussed further in the next section.

Inter-dependencies between criteria

In this section, we discuss inter-dependencies between criteria in the original framework. Inter-dependencies resulting from the addition of criteria or sub-criteria to the framework were discussed in the above section. Whenever possible without limiting framework scope, the framework was modified to eliminate identified inter-dependencies. Where such action would compromise the comprehensive coverage of the framework, the inter-dependencies were acknowledged rather than removed.

Syntactic and Semantic Criteria

Inter-dependencies were identified (a) within the set of semantic criteria and (b) between semantic and syntactic criteria. As discussed in the 'Semantic criteria' section, the original semantic definitions, expressed in terms of states, were operationalized in terms of identifiable IS and external (e.g. real-world) phenomena. As a result, *correct* was initially defined as having attribute values match property values for each represented external (e.g. real-world) instance. However, this resulted in inter-dependencies with other semantic criteria since a mismatch in key (i.e. identifying) attribute values could further lead to *ambiguous* (i.e. one identifiable IS data unit maps to multiple different external phenomena), *meaningless* (i.e. one identifiable IS data unit does not map to any external phenomena), or *redundant* mappings (i.e. multiple different identifiable IS data units map to the same external phenomenon) that violate the *unambiguous, meaningful,* or *non-redundant* semantic criteria, respectively.

The solution is to define two separate semantic correctness criteria, *phenomenon-correct* and *property-correct.* The first correctness criterion *phenomenon-correct* relates to the correctness of mapping identifiable IS data units to external phenomena. A violation would involve an unambiguous, meaningful, non-redundant mapping (based on key attributes) of an identifiable data unit to the *wrong* external phenomenon.

The second correctness criterion *property-correct* involves an identifiable data unit that maps correctly to the represented external phenomenon but has an incorrect representation of one or more non-identifier external properties by non-key attributes (i.e. un-matched values). To illustrate, an example of phenomenon-level correctness is when the ID field for a given employee record correctly maps to the real-world employee with that ID; whereas property-level correctness is when the recorded salary value matches the employee's actual salary.

It should be noted that although inter-dependencies are reduced and criteria definitions clarified by this framework revision, inter-dependencies between semantic criteria are not completely eliminated, since *property-correct* implies *phenomenon-correct,* which, in turn, implies a *meaningful* mapping. In fact, this inter-dependency originates directly from Wand and Wang's (1996) initial set of criteria where a *correct* mapping implies further that the mapping is *meaningful.* However, we consider the distinctions between these different cases significant (e.g. for error source analysis) and thus acknowledge rather than remove the inter-dependencies. A further concern is the apparent inter-dependency between the newly introduced semantic criterion *property-correct* and the syntactic criterion *conforming to integrity rules.* Incorrect property representation can result from either an illegal or a legal but invalid (i.e. incorrect, unmatched) attribute value. As currently defined, the former case seems to violate both the above-mentioned criteria; whereas the latter case seems to violate only the semantic criterion. However, it is possible that an IS attribute value may violate a syntactic formatting rule but still be able to be matched correctly to the relevant external (e.g. real-world) property value. We therefore clarify that *property-correct* is with respect to fidelity to external property values, but not necessarily to all specified integrity rules.

Based on this revision to the original semantic criterion *correct,* we then re-visit the issue raised in 'Missing criteria or sub-criteria' of possible sub-criteria. The discussion and examples given in that section apply without amendment to the addition of the *present, appropriate,* and *matching* sub-criteria to the new criterion *property-correct.* However, these sub-criteria do not apply to the new criterion *phenomenon-correct,* since we have said that any violation of this criterion is, by definition, an unambiguous, meaningful, and non-redundant mapping (therefore necessarily involving a *present* and *appropriate* but *non-matching* key value) or it would violate one of these other three semantic criteria instead. In other words, the key value successfully identifies exactly one external phenomenon (implying that the key value exists and is of the correct type) but it is the *wrong* phenomenon (implying that the key value does not match the identifier for the represented external phenomenon). So the data unit actually represents a different external phenomenon than that identified by the data unit's key values. We therefore conclude that the three new sub-criteria should be added to the *property-correct* but not the *phenomenon-correct* criterion.

Pragmatic criteria

Inter-dependencies were identified between pragmatic criteria relating to data delivery, in that information must first be *accessible* to judge whether it is *understandable* (and other presentation aspects) and must be *understandable* before judging whether it is *suitably* and *flexibly presented*. Conversely, information presentation affects perceived understandability and accessibility. In this case, we judged that, although inter-dependent, these criteria each represented essential and distinct quality aspects whose removal would result in a less comprehensive coverage of information quality. However, the sub-dimension *timely* was removed from *suitably presented* and made a separate delivery-related criterion. This restricts the criterion *suitably presented* to presentation style aspects (i.e. layout, precision, units), thus simplifying and clarifying its semantics. Further, it serves to acknowledge the critical importance of timeliness as a quality aspect in its own right, an issue raised in both academic and practitioner focus groups.

Further inter-dependencies between *understandable* and many other criteria (beyond those relating to data delivery) were identified. Essentially, information must be understood before its relevance, value, and perceived syntactic and semantic quality aspects can be judged. After consideration, the best response was judged to be explicit acknowledgement of the inter-dependency.

Finally, we consider the inter-dependencies between the pragmatic criteria *valuable* and most other criteria (insofar that satisfying other quality criteria implies high value), and especially with the pragmatic criteria *relevant*. Although these inter-dependencies were explicitly acknowledged; *valuable* was initially retained as a placeholder for domain-specific quality criteria that might not have been covered elsewhere in the framework. Focus group discussion failed to elicit any examples of such domain-specific criteria that did not fit into the framework (assuming the framework is revised as discussed in the section on 'Missing criteria or sub-criteria'), even though representatives of both general business and specialized technical applications (i.e. geographic information systems) were included in the focus groups. Furthermore, the evident confusion introduced as a result of these acknowledged inter-dependencies became clear during the course of the focus groups. The feedback clearly indicated that participants felt that the concept of *valuable* was too general and abstract to ensure consistent interpretation (i.e. rather it was likely to be understood quite differently by different people) or to convey any meaningful information. It was therefore judged not to be useful as a specific quality criterion and removed from the framework.

In a related issue, focus group feedback highlighted the fact that the *sufficiency* aspect of the original criterion *valuable* should instead be considered in the criterion *relevant* with respect to the types of information available. This aspect of quality is not considered elsewhere in the framework. Therefore, the criterion *relevant* can be replaced with the more comprehensive criterion *type-sufficient,* defined as the degree to which the given data set includes all of the types of information (i.e. *data intent)* useful for the intended information use. This is discussed further in the next section.

Criteria definition

In this section, we discuss focus group feedback relating to identified ambiguities in criteria semantics or wording not caused by dependencies between criteria (discussed in the above section). With respect to criteria semantics, it was evident from all of the focus groups that participants regarded the presence of redundancy in a data collection as quite common and not necessarily an indication of poor quality. In fact, they referred to replication, a synonym for redundancy with positive rather than negative connotations, as an integral part of effective organizational data management.

The argument presented in the 'Semantic criteria' section was that both *meaningless* and *redundant* data represented a potential rather than a definite quality problem and therefore should be treated similarly. However, as a result of the focus group feedback, the two cases could be clearly differentiated in that only the latter might be deliberately introduced because of associated benefits (e.g. with respect to improved access time for geographically dispersed consumers). The response to this observation is to redefine the quality criterion *non-redundant* as *consistent,* that is, not having duplicates or having acceptably consistent duplicates. Acceptable consistency is defined as either having consistent replicates (i.e. with matched attribute values) or inconsistency that is resolved within a time frame acceptable in the context of replicate use.

Considerations related to the impact of privacy laws on information quality led to the elaboration of the original definition of the pragmatic criterion *security* as 'appropriately protected from damage or abuse (including unauthorized access)' to include unauthorized use or distribution.

Another source of confusion raised by end-users was their difficulty in distinguishing between *suitably presented* and *relevant* (or the substituted *type-sufficient),* End-users' understanding of the type of information (i.e. attribute or field types) available is commonly based on what is displayed or made available through the presentation interface. Therefore, they tended to view this issue as just another sub-criterion relating to presentation rather than a separate and independent criterion. After further consideration, neither the original criterion *relevant* nor the newly proposed criterion *type-sufficient* are included in the revised framework. Instead, the sub-criteria *includes suitable field types* and *the selection of displayed field types easily changed* are added to the existing sub-criteria of *suitably presented* and *flexibly presented* respectively.

Other identified ambiguities in criteria definition are related to wording. For example, the term *meaningful* was often misinterpreted as important or significant rather than as defined in terms of a mapping cardinality constraint. Thus the implicit connotations of the English word took precedence over the definition given. Therefore the names of all the semantic criteria were amended to include explicit references to mapping, for example, *mapped meaningfully, mapped completely,* etc.

Framework context

Discussions relating to framework context helped to further clarify the scope and boundaries of the research.

Specialized data types

Focus group participants raised questions regarding whether specialized application domains, such as scientific data, were addressed by the framework, with spatial applications such as geographic information systems given as a specific example. This was discussed in the section on 'Interdependencies between criteria' in the context of the criterion *valuable,* which was deleted from the framework following the failure to identify any domain-specific criteria not covered at least generally by other framework criteria and the acknowledged confusion caused by the criterion's inherent inter-dependencies and ambiguity. It was additionally observed that although the framework did encompass spatial quality criteria, it was at a level that might potentially be too general to be useful in the context of specialized spatial applications. Therefore, the decision was made to explicitly acknowledge that the framework targeted (i.e. was specifically developed for) general business applications, although it might provide useful guidelines (i.e. a starting point) for conceptualizing quality even in specialized application domains. That is, if any domain-specific criteria exist in specialized application areas; it was judged more effective that individual organizations add them explicitly to create variants of the basic framework.

Unit of analysis

Questions were raised regarding the framework's intended unit of analysis, specifically whether it targeted data sets or individual data attributes (i.e. columns in the relational context). In the context of common organizational quality assessment requirements and the framework's potential for supporting those requirements, some practitioners felt that it was important to be able to assess not only entire data sets (e.g. the customer information relation) but also individual relational columns (e.g. the address column from that relation). On reflection, we realized that this represented one example of a more general issue. The general issue is that of quality assessment for data sets that do not include identifiers (e.g. any set of non-key columns in the relational context). In such a case, the individual data units (e.g. non-key field values from a record in the relational context) comprising the data set cannot be mapped to specific external phenomena. Two questions arise consequently: can the current framework support this type of assessment and how important is it to support this type of assessment?

It is immediately evident that because semantic category criteria are based on IS/real-world mappings, their evaluation requires identifiable data units in order to establish the necessary correspondence between data and external phenomena. Therefore, although parts of the framework are still relevant; the framework as a whole cannot support such quality assessments.

On first glance, it appears that such assessments are critically important to answer questions such as: *how reliable is stored customer address information?* However, closer examination reveals that this question is directed against the customer address attribute with respect to the customer identifier attribute(s), that is, *when we retrieve the address for any given customer, is it reliable?* If an address retrieved for a given customer is *reliable,* that means it is the correct address for that particular customer. Thus, such questions still involve data sets with identifiers, that is, identifiable data units that can each be mapped to individual external phenomena. In fact, the only type of quality assessment that is directed against an individual non-key attribute or attributes in isolation would be in the context of aggregation tasks, for example, *If we calculate the* *average or total employee salary, is it reliable?* In this case, there is no need to map salaries to employees. However, it is relatively rare that individual attributes are assessed for quality *only* with respect to aggregation tasks.

Objective versus subjective quality contexts

Several questions raised during focus group discussions highlighted contextual differences between the objective and subjective components of the framework in terms of the types of data and metadata that can be considered in practice (e.g. in quality assessments based on the criteria defined in the framework).

Although the syntactic criterion can be defined theoretically as *conformance to metadata (i.e. data integrity rules),* in practice, actual conformance assessments at the syntactic level would be objectively judged against existing database integrity rules as they are the only integrity rules explicitly specified and practically accessible. However, information consumers generally do not know which integrity rules have been specified; therefore, subjective consumer judgments of perceived conformance at the pragmatic level would be in the context of their own understanding of the applicable integrity constraints.

Similarly, objective quality criteria can be practically assessed only with respect to derived data that are stored; whereas assessment of subjective quality criteria necessarily includes both derived data that are stored and that which are calculated, since consumers would not normally be able to distinguish between the two cases.

In fact, as long as such differences between objective and subjective quality perspectives are explicitly acknowledged and understood in using the framework, they represent one of the potential strengths of the framework, as discussed in the third section. To reiterate, comparisons between objective and subjective quality assessments can be used to check for discrepancies that are likely to signify a quality problem (and that may not be immediately obvious from only one type of assessment) and may facilitate analysis into the source of the quality problem. For example, differences between syntactic and perceived syntactic quality assessments may be due to significant omissions in the integrity rules specified in the initial schema (i.e. data model problems).

Revised framework and comparison

As a result of focus group feedback, the scope of the semiotic information quality framework discussed in this paper can be clarified as follows. The framework is specifically intended for general business applications with structured data and for use with data sets that include identifiers (i.e. key attributes) allowing data units to be mapped to external (e.g. real-world) phenomena and vice versa.

Focus group feedback was also used as a basis for refining the original quality criteria, especially for the pragmatic category. The revised set of quality criteria and their definitions for each quality category is shown in Table 24.3, with any sub-criteria listed in parenthesis after the criterion name. Note that the terms *external phenomenon* and *phenomena* refer to external (e.g. real-world) instances.

Table 24.3 Revised quality criteria by category

Syntactic Criteria (based on rule conformance)

Conforming to metadata, i.e. data integrity rules: Data follows specified data integrity rules

Semantic Criteria (based on external correspondence)

Mapped completely: Every external phenomenon is represented

Mapped unambiguously: Each identifiable data unit represents at most one specific external phenomenon

Phenomena mapped correctly: Each identifiable data unit maps to the correct external phenomenon

Properties mapped correctly (present, appropriate, matching): Non-identifying (i.e. non-key) attribute values in an identifiable data unit match the property values for the represented external phenomenon

Mapped consistently: Each external phenomenon is either represented by at most one identifiable data unit or by multiple but consistent identifiable units or by multiple identifiable units whose inconsistencies are resolved within an acceptable time frame

Mapped meaningfully: Each identifiable data unit represents at least one specific external phenomenon

Pragmatic criteria (use-based consumer perspective)

Accessible (easy, quick): Data are easy and quick to retrieve

Suitably presented (suitably formatted, precise, and measured in units; includes suitable field types): Data are presented in a manner appropriate for their use, with respect to format, precision, units, and the type of information displayed

Flexibly presented (easily aggregated; format, precision, and units easily converted; the selection of displayed field types easily changed): Data can be easily manipulated and the presentation customized as needed, with respect to aggregating data and changing the data format, precision, units, or type of information displayed

Timely: The currency (age) of data is appropriate to their use

Understandable: Data are presented in an intelligible manner

Secure: Data are appropriately protected from damage or abuse (including unauthorized access, use, or distribution)

Allowing access to relevant metadata: Appropriate metadata are available to define, constrain, and document data

Perceptions of the syntactic and semantic criteria defined earlier

The revised quality criteria in Table 24.3 can be compared with the initial list of quality criteria in Table 24.2 to identify criteria whose definition or use in the framework were most affected by the empirical refinement process, as illustrated in Table 24.4. Other revisions relate to terminology or to framework context and thus are not shown in Table 24.4.

The revised semiotic information quality framework can then be compared to other information quality frameworks proposed previously. There have been a number of proposals that focus on a specific application domain (e.g. web quality from Barnes and Vidgen, 2002) or that consider information quality indirectly as one factor in a broader IS perspective (e.g. in the context of measuring IS success in DeLone and McLean, 2003 or modeling IS systems in Ballou *et al.,* 1998). Although these proposals are subject to some of the same criticisms relating to rigor and consistency discussed in the introductory section we restrict our comparison here to those information quality frameworks that are generic (i.e. not focused on a specific domain), applicable to general business applications (i.e. suitable for structured data in business databases or data warehouses), and frequently mentioned in recent information quality literature (i.e. from the last decade). Frameworks can be compared based on a number of different considerations. For example, Eppler's (2001) survey evaluates the clarity, positioning, consistency, conciseness, and practicality (in terms of examples and tools) of information quality frameworks. Our comparison in Table 24.5 focuses

Table 24.4 Summary of major revisions of quality criteria by category

Quality category	**Initial quality criterion affected**	**Revision**
Syntactic	None	
Semantic	*Correct*	Sub-divided into *property*- and *phenomenon-correctness* with further sub-criteria *present, appropriate, matching* added to the former
	Non-redundant	Re-defined in terms of consistency
Pragmatic	*Suitably presented*	Sub-criterion *timely* promoted to separate criterion
	Valuable	Eliminated
	Relevant	Definition changed to type-sufficient and demoted to sub-criteria of *suitably / flexibly presented*
	Secure	Definition amended to include use and distribution issues
		Addition of new criterion, *including access to metadata*

Table 24.5 Comparison of frameworks

Consideration	**English (1999)**	**Wang and Strong (1996), Kahn et al. (1997), Kahn et al. (2002), Lee et al. (2002)**	**Wand and Wang (1996)**	**Redman (1996)**	**InfoQual**
Derivation and definition of categories	*Ad hoc*	Empirical (original paper), Ad-hoc (follow-up papers, PSP/IQ model)	*Ad hoc*	Logical	Theoretical (semiotics)
Selection of criteria derivation method	Not discussed explicitly for framework	Based on stated assumption that empirical method best represents consumer perceptions	Theoretical (ontological view of IS)	Not discussed	Theoretical (semiotics)
Derivation and definition of criteria	*Ad hoc*	Empirical	Theoretical (mapping cardinality)	*Ad hoc*	Objective criteria: theoretical (integrity theory/mapping cardinality) Subjective criteria: literature-based (initial) and empirical (refined)
Classification of criteria into categories	*Ad hoc*	*Ad hoc* (initial) and empirical (refined)	Not applicable	*Ad hoc*	Theoretical (automatic consequence of selecting criteria derivation method)
Inter-dependencies considered?	No	No	No	Yes	Yes

(*continued*)

Table 24.5 Continued

Consideration	**English (1999)**	**Wang and Strong (1996), Kahn et al. (1997), Kahn et al. (2002), Lee et al. (2002)**	**Wand and Wang (1996)**	**Redman (1996)**	**InfoQual**
Criteria coverage (compared to InfoQual)	*Missing:* details regarding reliability, *security* criterion, *access to metadata* criterion, and *pragmatic* criteria based on perceptions of *syntactic and semantic* criteria.	*Missing: syntactic* criterion, details regarding reliability, *access to metadata* criterion, *pragmatic* criteria based on perceptions of *syntactic and semantic* criteria	*Missing:* does not consider *syntactic* and *pragmatic* categories of criteria.	*Missing:* details regarding reliability, *security* criterion, and *pragmatic* criteria based on perceptions of *syntactic and semantic* criteria. *Additional:* some consideration of data model and data storage quality	
Category and classification consistency	Inconsistency in classification of criteria into categories	Inconsistency in category names, definitions, and use in criteria classification	No classification done, categories consistent	Some inconsistency in classification	Theoretical foundation for these steps ensures consistency

on differences in framework development – in terms of the research approach(s) adopted and consideration of inter-dependencies – and the resultant implications for framework scope (i.e. specific criteria coverage) and consistency.

The most obvious difference between the frameworks highlighted by the table is the difference in research approach adopted. Only InfoQual provides a consistent theoretical basis for all of the development steps – with the single exception of the derivation of subjective quality criteria which is intrinsically dependent on information consumer judgements and thus requires empirical feedback (or industrial experience) to ensure relevance. Wand and Wang (1996) provide a rigorous basis for deriving and defining objective criteria using a theoretical approach, but are limited in scope and still rely on an *ad hoc* derivation of quality categories. Redman's derivation and definition of categories is termed *logical* rather than *theoretical* because although it is a result of logical reasoning and clearly stated objectives, it is not based on a systematic theory. Finally, neither English (1996) nor the frameworks based on the same set of empirically derived criteria (Wang and Strong, 1996; Kahn *et al.,* 1997, 2002; Lee *et al.,* 2002) provide any theoretical basis for their frameworks. For convenience, we will refer to the latter set of frameworks as *Wang's frameworks*.

The consequence of the lack of theoretical basis is clearly demonstrated when framework consistency is evaluated, especially with respect to the classification of criteria in categories. With the exception of InfoQual, all of the multi-category frameworks exhibit inconsistency in criteria classification and Wang's frameworks further show inconsistency (and ambiguity) in category definition.

Although the quality categories were empirically derived in Wang and Strong's (1996) original paper, the subsequent papers defined a new set of *ad hoc* quality categories (termed the *PSP/IQ* model) and re-classified criteria based on these new categories. The limited semantic basis for the selection of quality categories and their use in classifying the quality criteria in these frameworks is clear from both (1) the substantial changes evident in category names, definitions, and member criteria in successive papers (Wang and Strong, 1996; Kahn *et al.,* 1997, 2002) and (2) naming and definition ambiguities in categories and criteria, resulting in the lack of clear semantic differentiation between different categories or between a category and its criteria. Examples of the latter case include the *dependable* and *sound* categories in Kahn *et al.* (2002) and Lee *et al.* (2002); the *useful* and *effective* categories in Kahn *et al.* (1997), or the *access* category and its *accessible* criterion in Wang and

Strong (1996). Classification inconsistencies include the inclusion of the *believable* and *reputation* criteria in the *usable* rather than the *sound* or *dependable* categories, where intuitively it would be expected that these criteria are more directly related to reliability than usability.

Classification inconsistencies can also be clearly observed in English (1999) based on the specified category and criteria definitions. For example, although *precision* and *accessibility* are explicitly defined as being dependent on data use, they are classified as being *inherent* – a quality category explicitly defined by English as use-independent.

Redman's (1996) classification of criteria also shows inconsistencies. The scope of his framework includes data model quality (i.e. relating to the quality of *metadata* such as conceptual views) and data storage quality (i.e. relating to the quality of data *representation*) as well as information quality (i.e. relating to the quality of data *values* – both stored and received); with some ambiguity introduced as to the classification of criteria between these categories. For instance, schematic *conceptual view quality* is considered a separate category. It includes not only criteria relating to data model quality such as the *naturalness* and *clarity* of the entities and attributes defined but also criteria relating to information quality such as *accessibility* of data values. Similarly, the definitions of *format suitability* and *format flexibility* criteria in Redman's *data representation* category include both storage aspects (e.g. suitability/flexibility for specific/different physical media) and presentation aspects (e.g. suitability/flexibility for specific/different users), where the latter clearly relate to subjective views of information quality rather than to data storage or representation quality.

With respect to consideration of inter-dependencies between proposed criteria, Eppler (2001) notes that it is rarely considered in information quality frameworks proposed to date despite its importance for understanding the semantics and practical implications of an information quality framework. To illustrate the potential significance of such inter-dependencies, consider their acknowledged impact on the choice of appropriate analytic methods to be used, for example, in the application of such a framework to instrument development (see Straub *et al.*, 2004) for subjective information quality assessment.

Of the frameworks considered, only Redman's (1996) and InfoQual include any consideration of inter-dependencies between proposed criteria. Examples of significant inter-dependencies in Wang's frameworks that are not explicitly acknowledged or justified include those between *believability* and *reputation* criteria and between *ease-of-understanding* and *interpretability* criteria. Similarly, in English (1999), most of the criteria included in his *inherent* quality category (defined as use-independent) contribute to the criterion of *rightness* from his *pragmatic* quality category (defined as use-dependent). As discussed earlier in the section on 'Inter-dependencies between criteria', even Wand and Wang's (1996) restricted and theoretically-derived framework has an unacknowledged inter-dependency between the *correct* and *meaningful* criteria, where the former implies the latter.

When comparing the coverage (i.e. scope) of the different frameworks, we consider only significant omissions or additions with respect to the quality criteria defined by InfoQual.[2] Only InfoQual clearly differentiates between objective criteria and subjective perceptions of those criteria. Only Wand and Wang (1996) and InfoQual use mapping errors as the basis for deriving quality criteria relating to reliability (also called *accuracy, correctness,* etc.) and thus consider details of *reliability* in terms of the specific mapping cardinalities (*unambiguous, meaningful,* etc.). Coverage of InfoQual's syntactic category criterion, *security* criterion, and *access to metadata* criterion are inconsistent across the frameworks. As discussed earlier, Wand and Wang do not consider syntactic or pragmatic category criteria at all.

Notably, only Redman's framework contains criteria not considered by InfoQual, although these relate to data model or data storage quality rather than information quality. Data model quality has been more comprehensively treated by other authors (Krogstie *et al.,* 1995; Wand and Weber, 1995; Krogstie, 2001); however, data storage quality has not received the same attention in the literature. For instance, Redman describes criteria relating to the storage format's appropriateness or portability for different recording media (in effect, for different physical instances of the data). Such considerations can be important for an organization with enterprise information systems that may include multiple copies of data stored on different types of physical media. In the semiotic context, such criteria relate to the syntactic level as both *stored data format* and *stored data physical instances* (on physical media) can be considered signs (i.e. nested signs for the hierarchically structured levels of IS internal storage representation as described in the third section). Thus, a possible new adaptation or extension of InfoQual to include data storage quality considerations such as storage format quality is compatible with and naturally supported by the framework's existing theoretical foundation in semiotic theory.

Finally, we note that Redman considers both inter-record (i.e. replication) and intra-record sources of redundancy, where the latter case is the result of record fields with overlapping semantics. An example is a record containing both postal or zip code and state, where state information is redundantly determined by the code. Correctness of fields within a record (regardless of whether they overlap) is described by the *properties mapped correctly* criteria in InfoQual. However, in this case, no update lag should be allowed – the fields must be updated together to ensure consistency and thus correctness.

Conclusion

In summary, the comparative analysis from the previous section clearly shows that all of the frameworks except InfoQual suffer from limitations with respect to consistency and/or coverage. InfoQual addresses these problems by providing a consistent theoretical foundation for (1) the derivation and definition of quality categories, (2) the selection of derivation methods for quality criteria and consequent automatic classification of criteria into categories, (3) the derivation of objective quality criteria, and (4) the integration of objective, theoretically-based and subjective, non-theoretically based views of information quality. The use of empirical feedback to refine the framework ensures its relevance, especially with respect to the subjective quality view. The utility and power of using semiotic theory as the underlying theoretical foundation for the framework is further demonstrated by its relevance to unanticipated and new applications, for example, for data storage quality.

Since quality information is required for effective decision-making in an organization; continuous information quality management – including information quality assessment, problem identification and source analysis, and improvement strategies – is an essential element of decision support. A framework such as InfoQual that clearly and consistently defines the quality categories and criteria to be considered is an important pre-requisite for such a management program.

The explicit intention of the research reported here was to provide an information quality framework that could serve as a basis for further work in information quality in general and in information quality assessment in particular. Therefore, future work following on from this would include the development of assessment tools and techniques based on this framework. Additional areas of potential work include the evaluation of the utility and potential application of this framework to other aspects of information quality such as improvement and management and to specialized application contexts involving, for example, spatial or scientific data. Another possible direction would be to explore the application of InfoQual to data storage quality to include, for example, consideration of criteria related to storage format as outlined above.

Notes

1. We prefer the more inclusive term *external* to the frequently-used term *real-world* (e.g. in Wand and Wang, 1996), because of the latter's connotations that only concrete physical and not socially constructed phenomena (e.g. quotas) are considered.
2. Other apparent differences are shown not to be significant after a careful analysis of such factors as criteria overlap (i.e. what is the actual semantic coverage of additional criteria?) and validity (e.g. are the proposed criteria generic – applicable across application domains and data types?).

References

Ballou, D., Wang, R.Y., Pazer, H. and Tayi, G.K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality, *Journal of Management Science* **44**(4): 462–484.

Barnes, S.J. and Vidgen, R.T. (2002). An Integrative Approach to the Assessment of e-Commerce Quality, *Journal of Electronic Commerce Research* **3**(3): 114–127.

Barnouw, E. (ed.) (1989). *International Encyclopedia of Communications,* Oxford: Oxford University Press.

DeLone, W.H. and McLean, E.R. (2003). The DeLone and McLean Model of Information System Success: A ten-year update, *Journal of Management Information Systems* **19**(4): 9–30.

English, L. (1999). *Improving Data Warehouse and Business Information Quality,* New York: John Wiley & Sons, Inc.

Eppler, M.J. (2001). The Concept of Information Quality: An interdisciplinary evaluation of recent information quality frameworks, *Studies in Communication Sciences* **1:** 167–182.

Gendron, M. and Shanks, G. (2003). The Categorical Information Quality Framework (CIQF): A critical assessment and replication study, *in Proceedings of the Pacific-Asia Conference on Information Systems,* (Adelaide, Australia, 2003), Adelaide, South Australia: University of South Australia, 1–13.

Kahn, B.K., Strong, D.M. and Wang, R.Y. (1997). A Model for Delivering Quality Information as Product and Service, in *Proceedings of Conference on Information Quality,* (Massachusetts Institute of Technology, Cambridge, MA, USA, 1997), Cambridge, MA, USA: Massachussets Institute of Technology, 80–94.

Kahn, B.K., Strong, D.M. and Wang, R.Y. (2002). Information Quality Benchmarks: Product and service performance, *Communications of the ACM* **45**(4): 184–192.

Krogstie, J. (2001). A Semiotic Approach to Quality in Requirements Specifications, in: *Proceedings of IFIP 8.1 Working Conference on Organizational Semiotics,* (Montreal, Canada 2001), London: Chapman and Hall, 231–249.

Krogstie, J., Lindland, O.I. and Sindre, G. (1995). Defining Quality Aspects for Conceptual Models, in *Proceedings of IFIP8.1 working conference on Information Systems Concepts (ISCO3): Towards a consolidation of views,* (Marburg, Germany, 1995), Berlin: Springer, 216–231.

Krueger, R.A. (1994). *Focus Groups: A practical guide for research,* Thousand Oaks, CA: Sage.

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y. (2002). AIMQ: A methodology for information quality assessment, *Information and Management* **40:** 133–146.

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

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

Price, R. and Shanks, G. (2004). A Semiotic Information Quality Framework, in *Proceedings of the IFIP International Conference on Decision Support Systems (DSS2004),* (Prato, Italy, 2004), Melbourne, Victoria, Australia: Monash University, 658–672.

Price, R. and Shanks, G. (2005). Empirical Refinement of a Semiotic Information Quality Framework, in *Proceedings of Hawaii International Conference on System Sciences (HICSS38),* (Big Island, Hawaii, USA, 2005); Silver Spring, MD: IEEE Computer Society Press, 1–10.

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

Shanks, G. and Darke, P. (1998). Understanding Data Quality in Data Warehousing: A semiotic approach, in *Proceedings of the MIT Conference on Information Quality,* (Boston, MA, USA, 1998), Cambridge, MA, USA: Massachusetts Institute of Technology, 247–264.

Straub, D., Boudreau, M.C. and Gefen, D. (2004). Validation Guidelines of IS Positivist Research, *Communications of the Association for Information Systems* **13:** 380–426.

Stamper, R. (1991). The Semiotic Framework for Information Systems Research, in: Nissen H, Klein H and Hirschheim R (eds.) *Information Systems Research: Contemporary Approaches and Emergent Traditions,* Amsterdam: North-Holland.

Wand, Y. and Wang, R.Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM* **39**(11): 86–95.

Wand, Y. and Weber, R. (1995). On the Deep Structure of Information Systems, *Information Systems Journal* **5:** 203–223.

Wang, R.Y. and Strong, D.M. (1996). Beyond Accuracy: What data quality means to data consumers, *Journal of Management Information Systems* **12**(4): 5–34.