A Semiotic Information Quality Framework: Theoretical and Empirical Development

Rosanne J. Price[1,2]
Graeme Shanks[2]

[1]Clayton School of IT, Monash University, Australia;
[2]Department of Information Systems, University of Melbourne, Australia
Email: [rosanne.price@infotech.monash.edu.au, gshanks@unimelb.edu.au]

ABSTRACT

Good quality information is required for effective business operations and decision-making; however, fundamental questions remain regarding the definition of information quality and the specific criteria used to define and assess it. This report describes the development of an information quality framework intended to address these questions. The approach adopted is motivated by the observed limitations of other quality schemes proposed to date with respect to scope and consistency. Semiotic theory, the philosophical theory of signs, is used to support rigor without sacrificing scope. It provides a theoretical foundation both for the framework structure—quality categories and their criteria—and for integrating objective and subjective quality perspectives. Empirical methods are then used to refine the framework, especially the subjective components related to consumer quality perceptions. Specifically, practitioner and academic focus groups served to clarify the scope and boundaries of the research; to identify inter-dependencies, ambiguities, and gaps in the initial set of quality criteria; and to provide guidance for any future work in quality assessment and management based on the framework.

Keywords: information quality, data quality, semiotics

1. INTRODUCTION

The effectiveness of an organization is dependent on the quality of its information, which can only be ensured through continuous quality assessment and management. However, this pre-supposes agreement as to the definition of information quality (IQ) and the specific criteria against which such quality should be assessed. In fact, this currently represents an area of active debate and research. Addressing these research issues is an important step in establishing a basis both for developing IQ 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. Objective measures of IQ 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 considerably 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 report) with respect to actual user requirements, leading to consumer perceptions of poor information quality. Furthermore, even if data meets 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 IQ 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 and what is retrieved from data collections. In this report, 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 has been delivered to, presented to, and interpreted by the user. Thus the term *information quality (IQ)* refers to both objective views of stored data quality and subjective views of received data quality. As described earlier in this section, IQ research can then be characterized based on the view(s) of quality considered.

IQ research is further characterized by the range of research approaches employed, i.e. empirical, intuitive, or theoretical. In addition, a literature-based approach, based on review and analysis of existing quality literature, is often used to support one of the other three approaches or to survey and compare existing quality frameworks such as that found in [Eppler, 2001]. We note further that the focus here is on research specifically directed at IQ but generic with respect to application domain, rather than work specific to a particular application domain (e.g. [Barnes and Vidgen, 2001; Chengular-Smith et al, 1999]) or that consider IQ only indirectly in a broader IS context (e.g. [Ballou et al, 1998; DeLone and McLean, 2003]).

Empirical research approaches such as that in [Wang and Strong, 1996; Kahn et. al., 1997; Kahn et. al., 2002; Lee et al., 2002] rely on information consumer feedback to derive quality criteria. This has important implications for the quality criteria defined in this manner: because they are based on information consumer feedback rather than on a systematic theory, there are likely to be some inconsistencies, redundancy, and/or omissions. In particular, and as previously noted also in [Eppler, 2001], the quality criteria defined in [Wang and Strong, 1996; Kahn et. al., 1997; Kahn et. al., 2002; Lee et al., 2002] have significant inter-dependencies that are not explicitly acknowledged or justified (e.g. *believability* and *ease-of-understanding* subsume *reputation* and *interpretability* respectively). Other criteria such as *objectivity* are not generic. That is, they do not apply across all domains (since some domains require subjective data) or across all data types (since some data types are subjective, e.g. recorded managerial rankings of goals by priority). Furthermore, the quality categories presented 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 subsequent papers [Kahn et. al., 1997; Kahn et. al., 2002;

Lee et al., 2002] following on from the original [Wang and Strong, 1996] and (2) from naming ambiguities (e.g. overlapping names such as the *useful* and *effective* categories in [Kahn, 1997]] or the *access* category and its *accessible* criterion in [Wang and Strong, 1996]).

Because the intuitive 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 [English, 1999] informal intuitive 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. For example, although *precision* is explicitly defined as being dependent on data use, it is classified as being *inherent*—defined earlier as use-independent. As with empirically-derived quality definitions, a large number of criteria inter-dependencies are observable (e.g. most of the *inherent* criteria contribute to the *pragmatic* criterion of *rightness*) and are not acknowledged or justified.

In contrast, theoretical approaches such as Wand's [Wand and Wang, 1996] evaluation of data 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 intuitive approaches. The drawback of this approach is with respect to scope. A purely theoretical approach to defining quality and quality criteria is necessarily limited in scope to objective product-based quality aspects, as acknowledged explicitly by the authors themselves. It is clear that a complete approach to defining quality must take into account suitability for a specific task from the consumer's perspective, i.e. *service* quality. This aspect of quality is necessarily subjective in nature, both with respect to establishing the relevant set of quality criteria to consider and with respect to assessing quality based on these criteria.

In summary, a review of existing IQ frameworks and their limitations motivates a different approach to defining IQ that maintains rigor, especially with respect to the definition of quality categories and classification of criteria into categories, without sacrificing scope, i.e. which incorporates both product and service quality perspectives in one coherent framework. The intention of this report is to provide a comprehensive discussion of both the theoretical and empirical stages of the development process for such a framework.

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.[1]

To ensure rigor, a theoretical approach was used wherever possible, i.e. in the first three steps. The first two steps were based on semiotics; whereas the third step employs database integrity theory and ontology. This raises the question of scope. To be comprehensive, an IQ framework must include subjective component(s). However, such components are not amenable to a purely theoretical approach for the reasons given earlier. 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 report is structured as follows. Section 2 reviews semiotic theory and its application in an Information Systems (IS) context. Section 3 presents a detailed exposition of the initial theoretical development of an IQ framework, including the definition of information categories based on semiotic theory and derivation of specific criteria for each category. We report on empirical feedback from focus groups in Section 4 and present the resulting refined framework in Section 5. Finally, Section 6 describes conclusions and future work.

2. A SEMIOTIC APPROACH TO INFORMATION QUALITY

Although semiotics has many different branches, the one most relevant to IQ is that proposed by Charles Pierce [Pierce, 1931-1935] and later developed by Charles Morris [Morris, 1938]. In particular, Morris describes the study of signs in terms of their logical components [Intl. Enc., 1989]. These are the sign's actual *representation*; its *referent* or intended meaning (i.e. 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'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 in terms of three semiotic levels: *syntactic* (between sign representations), *semantic* (between a representation and its referent), and *pragmatic* (between the representation and the interpretation). 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 IQ can be clarified and the applicability of semiotics to the formal definition of IQ justified.

A datum is maintained in a database or data warehouse precisely because it is representative of some external (e.g. real-world) 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, 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

[1] 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.

they 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 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 IQ [Shanks and Darke, 1998a; Shanks and Darke, 1998b]. Following Stamper's lead, the semiotic model adopted by these authors differs from ours in that they adopt additional semiotic levels beyond those proposed in classical semiotics. These include 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) levels. However, there is no theoretical foundation for these additional levels in semiotics. In fact, one could argue that the concepts described by these additional levels are already covered by the original three levels as follows. The social context of the social level 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.

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 IQ 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 IQ 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.

3. THE THEORETICAL FOUNDATIONS

In this section, we describe the theoretical development of an IQ framework intended to address the twin goals of rigor and scope in order to serve as a practical and sound basis for IQ research in general and assessment in particular.

In Section 3.1, we describe the quality categories defined based on the semiotic levels described in Section 2 and the approach used to derive criteria for each category. Section 3.2 describes the derivation of individual quality criteria for each category and Section 3.3. presents the initial list of quality criteria so derived.

3.1 DEFINING QUALITY CATEGORIES BASED ON SEMIOTIC LEVELS

In this section, we describe the structure of our proposed IQ 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 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 IQ criteria with its associated quality goals, evaluation technique, level of objectivity, and criteria derivation approach. In Section 3.1.1, we define basic terms and quality categories based on semiotic theory. In Section 3.1.2, we discuss information associated with each quality category relating to quality assessment and criteria derivation.

3.1.1 Definitions

Before discussing the semiotic levels and derived IQ 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 term *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 quotas) and allows us to make the distinction between data and meta-data.

In addition, references to *stored* data assume a single abstract IS representation. This representation subsumes the different internal IS levels of logical and physical storage representations and their transformations (e.g. logical files with records and fields, physical disk tracks with binary bits). In fact, these hierarchically structured representations can be considered nested signs describing the IS internals.

Definition 1. A *sign* is a physical manifestation (i.e. the representation) with implied propositional content (i.e. the referent) that may have 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 (i.e. the referent); and a human or machine interpretation involving some use of the datum (i.e. the interpretation). In this case, the intended meaning is the represented external phenomenon relevant to a specific application domain.

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) and a human or (usually) machine interpretation involving some use of the meta-datum (i.e. the

interpretation). In this case, the intended meaning is any “data about data”, including schema information, integrity constraints, or application-specific documentation. Thus, the referent is a represented rule constraining or documentation describing datum (represented external phenomena) or their organization (i.e. relevant to a specific data model).

Definition 4. *Data* is a set of datum collected based on their shared relevance to achieve some 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, definitions, 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 rules, definitions, and documentation relating to either the business application domain or to the underlying data model that form the IS design. Next, we present the definitions of each semiotic level and its derived IQ category.

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

Definition 7. The *syntactic quality category* describes the degree to which stored data conform to stored meta-data.

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

Definition 9. 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 10. The *pragmatic level* consists of any relation between a sign representation and its interpretation.

Definition 11. The *pragmatic quality category* describes the degree to which stored data that is delivered to information consumers (i.e. “received” data) is considered suitable and worthwhile for a given use, where the given use is specified by describing three components: an activity (i.e. one or more tasks), its context (e.g. the geographic location or organizational sub-unit where the activity occurs), and the characteristics of the information consumer (e.g. experience, knowledge, and organizational role).

Definitions 6, 8, and 10 for the semiotic levels and Definitions 7, 9, and 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

objective product-based and the pragmatic category to the subjective service-based quality views described in Section 1. In practice, the quality category definitions 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 focus particularly on the use-based aspect of sign interpretation described in Section 2. The pragmatic category definition effectively describes a relation between stored data, "received" data, and its use, where "received" data is the result of delivery, presentation, and construal of the stored data by the information consumer using the data (as defined earlier in Section 1). Thus, this represents an indirect relation between a sign representation and its interpretation. In the context of information quality, *use* is further described by an activity, its context, and user characteristics (see Definition 11); since any judgement regarding the suitability and worth of a data set are dependent on all of these 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 metadata (e.g. data integrity 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 data integrity rule has been defined, (2) a given employee record (or set of records based on foreign-key-based relational joins) that correctly represents a real employee (e.g. has matching details), and (3) employee information available in the database that is useful for the tasks performed by the information consumer.

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 Section 1. 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 (e.g. indicating a problem with initial system design or consumer expectations rather than with data design conformance or external correspondence.)

3.1.2 Information Associated with each Quality Category

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 relative level of objectivity in the evaluation of each quality category.

The ideal and operational quality goals for each quality category follow directly from the definitions of that category. In order to establish a practical basis for

syntactic quality assessment based on the operational goal, we assume that important requirements for conformance to definitions and documentation have been specified in terms of integrity rules (see also Section 3.2.1). In general, 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).

Table 1. Quality Category Information

	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 metadata	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[2]	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 (e.g. compare expected and perceived quality levels)
Degree of Objectivity in Evaluation	completely objective, independent of user or use	objective except for user determination of which specific external phenomena are relevant to an application and how they are mapped to data representations	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

[2] We assume that important requirements for conformance to definitions and documentation have been specified in terms of integrity rules.

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; however, since that is outside the scope of this report 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. For practical purposes, sampling techniques may be required to evaluate the level of conformance for large data sets.

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.

In terms of the relative objectivity in the evaluation 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 application relevancy and correspondence judgements be made. That is, decisions must be made regarding both precisely which set of external phenomena should be represented (i.e. are considered relevant to the applications which the data collection is intended to support) and the correct mapping of data to external phenomena and vice versa. This is important, for example, in judging the completeness and reliability of data. 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 as to which factors are of significant concern in their own use-based evaluation of information quality. That is, although an initial set of pragmatic quality criteria can be proposed based on an analytic review of quality literature, they require refinement and validation through

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

empirical research methods. In the next section, we discuss the derivation of quality criteria for each category in detail.

3.2 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, and
- quality 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, i.e. corresponding as closely as possible to common usage,
- quality 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 quality criteria should be comprehensive.

These are listed as goals rather than requirements since we cannot prove that the goals are satisfied—they can only be subjectively assessed over time through peer review and empirical feedback. We note that as a rule, quality schemes proposed in the literature do not explicitly consider or acknowledge inter-dependencies. However, it is our belief, as previously noted by Eppler (2001), that explicit recognition of inter-dependencies between quality criteria is an important pre-requisite for any framework intended to support quality assessment since the inter-dependencies may have implications for the analytic methods used in the evaluation. The derivation of syntactic, semantic, and pragmatic criteria are described in Sections 3.2.1 through 3.2.3 respectively.

3.2.1 Syntactic Quality Criteria

The single syntactic quality criterion is derived directly from the definition of the syntactic quality category based on database integrity theory. Note that although in the most general theoretical sense this would ideally be defined as conformance of data to metadata (including definitions, documentation, and rules, i.e. the data schema); this definition is operationalized as conformance to specified data integrity rules in order to serve as a practical basis for syntactic quality assessment. Essentially, this assumes that important requirements for conformance to definitions and documentation have been specified in terms of integrity rules (a more specific version of the earlier assumption relating to perfect metadata in Section 3.1.2).

1. *Conforming to Integrity Rules:*
   Data obeys the constraints described by the specified data integrity rules.

This quality criterion can be illustrated by the example given earlier in Section 3.1.1 for the syntactic quality category, i.e. no employee records have an age value more than 65, assuming such an integrity rule was specified.

3.2.2 Semantic Quality Criteria

The derivation of semantic quality criteria is based on the work by Wand and Wang [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 report 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, followed by derivation of quality criteria based on that analysis. This assumes an ontological view described in detail in [Wand and Weber, 1995] 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, there are no meaningless IS states, i.e. states 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. In contrast, our work is intended to serve as a basis for IQ assessment. To address the difference in goals, the original analysis and resulting list of derived quality criteria are amended as follows.

- Representation versus Interpretation Phase and Information System versus State-based 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]. Furthermore, this results in an emphasis on legal rather than actual (i.e. existing) real-world and IS states, where only the latter are practicable to assess.

In contrast, to support our goal of facilitating IQ assessment, it is more useful to analyse possible deficiencies from the interpretation phase in line with our emphasis on *information* rather than *data model* quality. We further shift the focus from the deficiencies of an information system (with all of its states and laws) to the deficiencies of a single information system state (ie. the data stored in the information system at a given time). Such a view is better suited (ie. provides better support) for evaluating information quality operationally based on the current database snapshot (ie. existing data). This ensures that the primary focus is on actual rather than legal states. 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, or meaningless data noted by Wand and Wang, amended as discussed in the points below.

- Reference to Real-World versus External Comparisons.

  Wand and Wang's paper is framed in terms of comparisons with the real-world based on Bunge's ontology [Wand and Weber, 1995], which limits consideration to concrete physical phenomena, excluding social or mental constructions (e.g. imaginary, socially constructed, supernatural phenomena). This is reflected in their terminology by the reference to "real-world" (i.e. concrete) phenomena. Other ontologies such as that described by Chisolm [Chisolm, 1996] consider both types of phenomena. In keeping with this broader perspective, we henceforth describe phenomena outside the IS using the more inclusive term *external* rather than *real-world* to support the broadest range of IS data types and application domains.

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

  Wand and Wang's definitions are expressed in terms of 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 IQ assessment. Therefore, we operationalize the definitions by using *identifiable IS data artefacts* and *external phenomena* (whose states can be sampled individually) instead of *database states* and *external* (e.g. real-world) *states* respectively. We use the term *identifiable data unit* to refer to the IS data artefact used to represent a single external phenomenon, where the term *identifiable* is used to indicate that the IS data artefact is uniquely identified in the data collection. For example, the identifiable data unit in the IS representing a single external phenomenon such as an employee might be a single relational record or a set of joined records in a relational database or an object in an object-oriented database.

  As discussed by Wand and Wang, these two perspectives are interchangeable when analysing data deficiencies, except in the special case

[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 report is on the quality of data—both stored (i.e. in-situ in the data collection) and received (i.e. retrieved as information by users).

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.

- Revision of the Definition of *Incomplete*

  Wand and Wang define *incomplete* as the case where a real-world state *cannot* be represented by the IS and ascribe this to an error in design. However, given our focus on *information* rather than *data model* quality and on *interpretation* rather than *representation* transformations, we amend the definition of *Incomplete* for our purposes to describe the situation when 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). An example of the latter case would be when a data entry clerk manually entering data into the IS accidentally omits an entry.

  Note that this implies that IQ 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 IQ versus data model quality discussed in the previous point.

- 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. Note that this issue is re-visited in Section 4.4 as a result of focus group feedback.

- Assumption of Perfect Analysis and Implementation.

  Wand and Wang 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 Wang claim that their derived criteria are completely use-independent. However, as discussed in Section 3.1.2 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. In practice, judgements as to the correct mapping of IS to external states and vice versa also have an element of subjectivity based on data use insofar as it shapes user perspectives.

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 phenomenon maps to at least one identifiable data unit.
2. *Unambiguous*:
   Each identifiable data unit maps to no more than one external phenomenon.
3. *Correct*:
   The mapping of external phenomena to identifiable data units is such that the reverse mapping preserves the original details of the external phenomena.
4. *Non-redundant*:
   Each external phenomenon maps to no more than one identifiable data unit.
5. *Meaningful:*
   Each identifiable data unit maps to at least one external phenomenon.

Note that 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 IS states as discussed earlier.

Together, the first 3 quality criteria express the minimal semantic quality requirement that each external phenomenon map to at least 1 identifiable data unit and each identifiable data unit map to no more than 1 external phenomenon. The full set of 5 criteria further restrict the mapping to be exactly 1-to-1, i.e. each external phenomenon maps to exactly 1 identifiable data unit and that each identifiable data unit maps to exactly 1 external phenomenon. This represents the optimal semantic quality requirement. To illustrate these 5 semantic quality criteria, we use the employee database introduced earlier. For simplicity, we assume that an employee is represented by a single database record. 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 at least one actual employee.

The semantic criteria and definitions listed above—especially the *correct* and *non-redundant* criteria—are considered further and refined based on the focus group results, as reported in Sections 4 and 5.

3.2.3 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, organizational context, and user background. That implies that the assessment of IQ using such criteria will be based on information consumer perceptions and judgements, since only they can assess the quality of the data relative to use. Furthermore, as explained earlier in Section 3.1.2, the selection of the pragmatic criteria is similarly subject to information consumer judgements as to which factors they consider important in their use-based quality assessments. 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 (discussed in Section 4). In this section, we discuss the initial list of pragmatic criteria and their rationale.

Before considering existing quality literature to construct an initial list of pragmatic 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 criterion. 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 is as follows:

1. Existing quality literature was reviewed for use-related criteria, where determination of use-related criteria was based on evaluation of criteria definitions or contextual descriptive text as presented in the literature for a given researcher. When there was no explanatory material given for a criterion, interpretation relied on common usage of the term based on other extant quality literature and dictionary definition. In addition to those criteria explicitly classified by an author as use-based, even those criteria not so classified by an author were considered for inclusion in the initial list of criteria. For example, *precision* was judged to be use-related in English [1999] based on the author's own definition even though he classified it as use-independent. The approach was to err on the side of inclusiveness for the initial list.
2. Examine the initial list of use-related criteria for synonyms, overlaps and inter-dependencies, or inconsistencies based on text analysis of the wording of the definitions or contextual descriptive text in quality literature or, as necessary, on dictionary definitions.
3. Identify any hierarchical groupings of criteria based on their semantics.

4. Compile a revised list with the goal of minimizing redundancy without sacrificing comprehensiveness.

As discussed earlier in Section 3.2, a distinguishing feature of the framework presented in this report is the consideration of inter-dependencies. Thus the goal during the initial selection and later empirical refinement of criteria was to identify in general any dependency between criteria. Subsequent comments during the reviewing process prompted a closer investigation of the identified inter-dependencies and their categorization as either *value* or *evaluation* dependencies, where one criterion's value or evaluation respectively is dependent on another criterion's value. For example, the perceived *understandability* of data (i.e. the value of the criterion *understandable*) is affected by the *suitability of* its *presentation*; however, the *suitability of* data *presentation* cannot even be evaluated unless that data is *understandable*. Thus the criterion *understandable* is value-dependent on *suitably presented,* while *suitably presented* is evaluation-dependent on *understandable.* In the rest of the report, all references to inter-dependencies between criteria refer to value-dependencies unless otherwise stated.

In step 3, it was observed that many of the specific criteria listed could be seen as specific aspects (sub-criteria) of more general quality criteria. In general, the association of sub-criteria to criteria was quite straightforward, e.g. *suitably formatted* and *suitably precise* with *suitably presented, easy to access* and *quick to access* with *accessible*, *easily re-formatted* with *flexibly presented*. 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 consists of thirteen criteria in total. As discussed above, six of these criteria describe subjective information consumer *perceptions* of (and are directly derived from) those objective quality criteria defined for the syntactic (*conforming to integrity rules)* and semantic (*complete, unambiguous, correct, non-redundant, meaningful*) categories. In combination, the five pragmatic criteria derived from the semantic category describe precisely the reliability of the data as a true and one-to-one mapping (i.e. representation) of external phenomena and thus subsume the commonly listed, inter-dependent, and loosely defined quality criteria *believable*, *reputable*, and *accurate* found in the literature. (Information judged to be reliable is necessarily believable, reputable, and accurate.) An additional seven use-based pragmatic criteria pertain either to the delivery (1-5) or to the importance (6-7) of retrieved data as follows.

1. *Accessible:*
   Data is easy and quick to retrieve.
   Sub-criteria: easy to access, quick to access.

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

3. *Flexibly Presented:*

Data can be easily manipulated and the data presentation customized as needed.
Sub-criteria: easy to aggregate, easy to change (i.e. convert) format, precision, or units.

4. *Understandable:*
   Data is presented in a manner easy to interpret.
5. *Secure:*
   Data is appropriately protected from damage or abuse (including unauthorized access).
6. *Relevant:*
   The types of data available (i.e. data intent) are pertinent to your use of this data.
7. *Valuable:*
   The data is useful and sufficient for (i.e. important for) your use of this data.

Criteria 1-5 above relate to information retrieval (1), presentation (2-3), comprehensibility (4), and protection (5) respectively. Whereas *suitably presented* describes whether the presentation of retrieved information is appropriate for its use, *flexibly presented* describes whether the presentation can easily be modified to suit different purposes. These two criteria represent groupings of related criteria commonly found in quality literature.

Note that, in the quality literature, the term *complete* is sometimes also used to mean that the types of data available (i.e. data intent) are sufficient for the use of the data. However, this definition overlaps with the definition of the criterion *relevant,* i.e. the types of data available are appropriate for the use of the data. Since *sufficiency* clearly implies *appropriateness*, we can see that one definition subsumes the other. Therefore, we employ the more restricted definition of the quality criterion *complete* (i.e. relating to the sufficiency of the data extent rather than the data intent). This has the twofold advantage of eliminating an inter-dependency (caused by semantic redundancy) between two quality criteria commonly observed in the quality literature and maintaining consistency with the definition of *complete* given for the semantic category. The definition of *relevant* is further considered and refined in Section 4.3.2 based on focus group results.

Finally, the 7th 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, it is included in the initial list of pragmatic criteria to act as a generic placeholder for those aspects of quality specific to a given application (i.e. domain-specific). 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. The problems and significance of this particular quality criterion has not, to our knowledge, previously been explored or acknowledged in the literature. This issue is re-visited in Section 4.3.2 as a result of focus group feedback.

The 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 regarding missing, incorrectly calculated, incorrectly addressed, duplicate, or spurious pay checks may result from problems with *incomplete* (i.e. missing) employee records, *ambiguous* employee identification numbers (i.e. matching more than one employee), *incorrect* addresses*, redundant* (i.e. duplicate) employee records*,* or *meaningless* employee records (i.e. not mapping to any employee) which are reflected in low consumer perceptions of 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, 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 specific spatial data depends on the nature of its lineage (i.e. the relative reliability of the specific data capture and transformation methods used to produce that spatial data). Thus, the general quality criterion *valuable* acts as a placeholder for the specific spatial quality criterion of lineage.

The pragmatic criteria described here are considerably revised based on focus group feedback, as reported in Sections 4 and 5.

3.3 SUMMARY TABLE OF THE INTIAL LIST OF QUALITY CRITERIA

Table 2 presents the initial list of quality criteria derived for each level using a theoretical research approach, with any sub-criteria listed in parentheses.

Table 2. Quality Criteria by Category

Syntactic Criteria (based on rule conformance)
Conforming to integrity rules

Semantic Criteria (based on external correspondence)
Complete, Unambiguous, Correct, Non-redundant, Meaningful

Pragmatic Criteria (use-based information 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

4. PRACTITIONER AND ACADEMIC 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 Section 3.1.2, i.e. that empirical techniques are required to solicit information consumer input as to the appropriate set of pragmatic quality criteria since by definition they relate to the subjective information consumer perspective. 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. 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. A secondary goal of the focus groups was to explore perceived potential benefits and implications of using the framework to support quality assessment or management, as judged by professionals concerned with data quality. 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.

Participants were asked to complete an individual opinion form evaluating the pragmatic criteria prior to their attendance at a two-hour focus group discussion of those criteria and of related quality issues. After the processing and analysis of the individual opinion forms and focus group discussions were complete, each participant received a follow-up report summarizing the opinion form and focus group feedback. We now describe each of these components of the study in more detail.

The intention of having participants complete individual opinion forms prior to the focus group discussion was twofold:

- to canvas individual opinions uninfluenced by other participants,
- to facilitate and guide the subsequent focus group discussion by:
    - familiarizing participants with the topics to be covered,
    - ensuring that participants gave some thought to the questions prior to the group discussion,
    - identifying in advance areas of misunderstanding, confusion, or lack of consensus that could warrant further explicit attention in the focus group for the purposes of clarification or exploration.

The individual opinion form (see Appendix 1) consisted of (1) an instruction section; (2) a section giving necessary background information relating to IQ research in general, the proposed IQ framework, the current research project and empirical study, and the form itself; and (3) the actual question section soliciting

participant responses. In the last section, participants were asked to evaluate the *clarity* (yes/no response), *independence* (yes/no response), and *importance* (using low/medium/high response) of each individual pragmatic category criterion by ticking the appropriate box and—at a minimum—writing explanatory comments for any *no* or *low* response. Participants were additionally asked about the comprehensiveness of the list of criteria. In particular, participants were asked whether *all significant aspects of data quality were included explicitly in the list* and, if not, to identify and define in their own words the missing criteria. The last question asked whether participants felt that the criteria *correct, meaningful, unambiguous,* and *non-redundant* should be combined into a single summary term *reliable.*

The intention of the focus group was both to confirm our understanding of the responses to and explore in more depth the basic questions asked in the individual questionnaire. Recommendations for conducting focus groups include selection of a fairly homogeneous group of between six and eight people [Krueger 1994]. On reflection, we judged that the people best placed to give a comprehensive assessment of the set of pragmatic quality criteria and their implications for quality assessment and management practices were IT practitioners or academics having direct responsibility for or research interest in information quality. The practitioner focus group consisted of seven IT practitioners including both data management/quality consultants and in-house IT professionals at varying levels of seniority. This included:

- a data warehouse report and application developer;
- two senior managers of data quality and information management units (one from a very large company, the other from a medium-size company); and
- four independent consultants specializing in data quality and in database or data warehouse design, where two of the consultants were founders and CEO/owners of consulting companies with the same specialization and another had had extensive experience with *TQM* (Total Quality Management).

The academic focus group consisted of seven academics whose research was in the area of data management, with particular sub-specializations (i.e. data management for…) as follows:

- three participants with expertise in decision support, one of whom had additional expertise in data mining and the other in data warehouses,
- two participants with expertise in conceptual modelling, one of whom specialized in spatial databases and applications,
- one participant with expertise in ubiquitous and distributed data mining and web-services, and
- one participant with expertise in scientific and transportation applications.

The plan for the focus group discussions (see Appendix 2) was then constructed based on the individual opinion form questions and feedback. All of the questions included in the individual opinion form were repeated in the focus group discussion. As a consequence of the individual form feedback, focus group participants were additionally asked whether the criterion *valuable* should be included to ensure comprehensive coverage of IQ characteristics. Specifically, they were asked whether they knew of any domain-specific criteria and—if so—whether the criterion *valuable* should be included as a general "catch-all" for any such domain-specific criteria. Finally, there was provision for additional general discussion of IQ in practice based on participant interests and experience. Based on his prior experience in conducting such focus groups, one of the researchers

(Graeme Shanks) was responsible for moderating the actual focus group discussions using the prepared plans as a guide. His mandate was to ensure that all of the plan's basic questions relating specifically to the list of pragmatic criteria were addressed; however, the extent of the discussion of each question varied widely depending on the different interests and concerns of each group's participants. The other researcher (Rosanne Price) was responsible for written notes recording the course of the discussion, and for prompting the moderator when issues needed further exploration. In addition, the focus group discussions were taped (using both analog and digital technology) and transcribed.

During the focus group discussion, participants were passionate about their views and experiences of quality issues and the challenges of ensuring quality. The wide-ranging discussion that ensued addressed topics such as defining, assessing, improving, and managing quality in organizations. Since the framework was presented as intended to be suitable 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.

The processing and analysis of the individual opinion forms and focus group discussions are discussed next. The individual opinion forms from the participants of a given focus group were initially analysed prior to the actual focus group discussion as follows:

1. for each question in the opinion form, the number of responses in each category was tallied (e.g. the number of *yes* responses for *clearly defined?* for each criterion; the number of responses indicating agreement, lack of agreement, or non-response respectively for criteria list comprehensiveness),
2. for each question in the opinion form, individual comments were collated, and
3. the resulting analysis of individual opinion form with tallies and collation was examined to determine whether the responses indicated a need for any additional questions to be asked explicitly in the focus group discussion (e.g. the confusion and lack of consensus evident from responses for the criterion *valuable* suggested that this issue should be raised explicitly and discussed further in the focus group discussions).

The steps involved in processing the focus group discussions were as follows:

1. the written notes, tape transcripts, and recorded tapes were compared for consistency and—for the transcripts—comprehensiveness,
2. intensive text analysis over the course of several weeks for each discussion group was conducted to identify the main discussion themes and categorize participant comments by theme,
3. the individual opinion form analysis was integrated with the focus group discussion analysis by (a) examining the collated opinion form comments to identify any themes not discussed in the focus group that should be additionally considered, and, since no additional themes were identified, (b) categorizing opinion form tallies and comments and collating them with focus group feedback using the existing list of focus group themes,
4. the resulting categorized and integrated opinion form and focus group feedback was then used to write a follow-up report for participants summarizing the feedback by theme, and
5. the follow-up report was sent to participants with a request for feedback regarding their opinions of the report itself.

After sending the follow-up report, the only two responses we received praised the objectiveness and/or comprehensiveness of the summary in reporting on the discussion. One of the IQ/database design consultants and company founders from the practitioner focus group noted that the report was "an accurate summary of what was said" and showed "evidence of objectivity and thoroughness in the way that you've recorded contributions that you judged to fall outside the framework". Another comment received was that "in my [the participant's] experience, much work in the data management and data quality field is compromised by unstated assumptions and agendas on the part of the researchers—I didn't get that view here".

In the rest of this section, we present the focus group results (based on both individual opinion forms and the group discussion) and our considered responses to the issues raised. In keeping with the intended goals outlined earlier, the focus group feedback served both to guide the refinement of framework criteria and to explore potential benefits and implications of applying the framework to IQ assessment or management. Discussed in Sections 4.1 through 4.4 respectively, focus group outcomes pertaining to framework refinement can be categorized as related to the framework context, missing criteria, inter-dependencies between criteria, or criteria definition. In these sections, our responses consist of an explanation of the framework refinement considered appropriate to address the specific concerns raised, with the resultant revised framework presented in Section 5. Discussed in Sections 4.5 through 4.6 respectively, focus group outcomes pertaining to perceived implications of framework use can be categorized as related to subjective quality assessment or objective quality assessment. Quotes from participants have been included where appropriate to illustrate and document our conclusions; however, participants are referred to by occupation rather than by name to maintain the anonymity required by governing ethics regulations. In these sections, our responses typically assess the significance of concerns raised or highlight possible actions that could be taken to address these concerns.

4.1. FRAMEWORK CONTEXT

Focus group discussions relating to framework context helped to further clarify the scope and boundaries of the research. The issues raised in these discussions were addressed as appropriate either by (1) explicit analysis and clarification of the issue's relevance with respect to the framework (e.g. see Section 4.1.2 below), (2) further explication of the framework definition to address the contextual issue (e.g. see Section 4.1.2 below), and/or (3) explicit acknowledgement that the issue is outside our current research scope (e.g. see Section 4.1.1 below).

4.1.1 Types of Data Considered

A fundamental question that emerged from focus group discussions related to the types of data or data collections addressed by the IQ framework. As discussed in detail below, collections of specialized data such as sampled, synthesized (interpolated), scientific (e.g. spatial data for geographic information systems), and non-electronic data sets may involve quality criteria distinct from those relevant to electronic data collections maintained for general business applications.

Where cost considerations require that a representative sample be separately maintained for a large data collection for analytic and other business purposes, the quality of this sample data set may be judged based on different criteria than that of the larger set, i.e. expected to be *representative* rather than

complete. One academic participant with experience in scientific applications repeatedly raised the issue of "a statistical sample…used for research or information (e.g. marketing) rather than processing uses of data. Statistical samples need representative data". The participant stressed that "often the quantity of data requires drawing a sub-set data set that is not necessarily complete but must be representative of the overall population of data to be 'correct'". Other participants in the group agreed, one specializing in data warehousing commenting that "companies are cost and time obsessed, so representative samples are important". Our view is that this is inherently an issue relating primarily to the quality of the sampling process rather than to the actual data set, a view supported by the observation that—in contrast to other quality criteria—a criterion such as *representative* cannot be judged against the data set alone, but only in comparison to the original data set from which it was drawn. However, the framework could be directly applied to a sampled data set by regarding it separately from the original collection and judging its completeness against the equivalent real-world population sample.

On consideration the other types of data discussed in the focus groups—synthesized, scientific, and non-electronic—were judged to be out of the current research scope.

Synthesized data refers to a special type of derived data which is based on interpolation from existing data, required, for example, to reconstruct or estimate missing data. The example given in the focus group by the academic participant with experience with transportation applications was that of "missing *journey to work* timings from a large transportation survey, so we had to estimate, reconstruct, or derive it on the basis of existing data". As with sampled data, the special characteristics of such data have potential implications for the criteria used to assess their quality. In this case, the quality of the synthesized data is related to their statistical properties, e.g. "whether synthesized data was within the *statistical range of probability*".

Other specialized data sets, such as scientific data, pose similar challenges. An example given by the academic participant specializing in spatial databases was that of spatial data quality concerns for geographic information systems. He felt that the proposed framework is generally consistent with these concerns, but at a level that may be too general to be useful in the specific context. This is discussed in more detail in the section on inter-dependencies between pragmatic criteria in Section 4.3.

Finally, questions relating to non-electronic data exposed the framework's underlying assumption that assessed data is electronically stored, especially with respect to presentation aspects of IQ (i.e. the pragmatic criteria *suitably presented, flexibly presented*). One of the consultants noted that the questions in the individual opinion form "assume that the data is all in the database, while in many cases it is still on paper or microfiche". Although the proposed framework may have much in common with and therefore serve as a guideline for specialized data quality, we therefore conclude that specialized domains may require some additional, different, or more specific quality criteria.

4.1.2 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). While one independent data quality consultant felt that "people using data think in terms of aggregates", another consultant felt that "data

quality depends on individual items [attributes or relational columns]" and thus that it was "important to consider the data quality of individual columns as well as whole databases". In the context of common organizational quality assessment requirements and the framework's potential for supporting those requirements, some participants thus felt that it was important to be able to assess not only entire data sets (e.g. customer information table) but also individual relational columns (e.g. the address column from that table). On reflection, we realized that this represented one example of a more general issue: quality assessment for data sets that do not include identifiers and thus cannot be mapped to the external world (e.g. any set of non-key columns in the relational context). 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/external mappings, their evaluation requires identifiable data units (i.e. record or set of records) 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 address attribute with respect to the customer identifier attribute(s), i.e. *When we retrieve the address for any given customer is it reliable?* Thus, such questions still involve data sets with identifiers, i.e. identifiable data units (consisting of one or more data items) that can each be mapped to individual external phenomena. Other quality assessments may be based on integrity rules involving one or more non-key attributes. Examples include constraints related to domain integrity (*e.g. are all salaries within the allowable range?*) or derived data (*is the calculated average or total employee salary reliable?*) In these cases, there is no need to map salaries to employees. This type of assessment can be handled using those parts of the framework (in the syntactic or pragmatic categories) that are relevant.

4.1.3 Objective vs. 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). For example, one of the consultants felt that "we should only speak of the data quality of base not derived values. It doesn't make sense to talk about the data quality of [derived] account balance data because it has to do with computational [process] not data quality". However, another consultant disagreed and felt that "since customers might not know that account balance was derived…they might still want to know how reliable it was". Similarly, with respect to metadata, the academic participant with conceptual modelling expertise asked whether "the syntactic level [of the framework] is applicable regardless of how integrity rules are implemented, in database integrity rules or application programming for example". To summarize, participants queried whether criteria applied only to data and metadata stored in the database.

On consideration, it is clear that it is only practical to assess objective quality criteria against data and metadata stored explicitly in the data collection; whereas the assessment of subjective quality criteria depends on consumer judgement and understanding, which may encompass data (e.g. derived data calculated on

demand) and metadata (e.g. assumed rules) that are not explicitly stored. For example, consider the syntactic criterion defined theoretically as *conformance to data integrity rules*. In practice, this criterion would be objectively assessed against existing (i.e. stored) database integrity rules at the syntactic level of the framework. However, at the pragmatic level of the framework, this criterion would be subjectively assessed (i.e. *perceived conformance*) against the set of integrity rules understood and considered applicable by individual consumers, since they generally do not know which integrity rules have been specified in the database. Similarly, objective quality criteria can be practically assessed only with respect to derived data that is stored; whereas assessment of subjective quality criteria necessarily includes both derived data that is stored and that which is calculated, since consumers would not normally be able to distinguish between the two cases.

More generally, participants contrasted objective quality assessments of the data (i.e. database state) and subjective quality assessments by consumers affected by factors such as data delivery, data presentation, intended data use, and consumer expectations. A consultant felt that it was important to begin with consumer assessments of information quality problems by "asking where does it hurt? where are the problems?". In contrast, a senior data quality and information manager felt although "these are good questions they are anecdotal". Instead, he countered that "baseline and continuous assessment of objective quality measures for core customer attributes critical to business success is essential to identify where the [quality] problem is and to justify expenditure on data quality projects". An academic specializing in decision support talked about "an experiment [he had conducted] where not just interface but also task resulted in very different data quality judgements". Other participants agreed, commenting that "you cannot separate the perception of data from that of the interface". Practitioners also highlighted the vulnerability of consumer quality perceptions both to the interface used to retrieve the data (i.e. data presentation), the intended use of the data, and the background of the information consumer (e.g. experience level). For example, one of the consultants in the practitioner focus group contrasted the different perception of "understandability to a person who has been there [in the company] 20 years versus someone who has been there 20 days". An academic participant commented in the individual opinion form that *understandability* "depends on the user background".

It is our contention that 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 Section 3.1.1. 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).

4.1.4 Data (state) vs. Information (process) vs. Data model quality

The difficulty of evaluating data quality (i.e. database state) in isolation was highlighted, especially in the case of subjective quality assessment. Apparent problems in data quality may actually be symptomatic of problems with the underlying data model or schema, with processes rather than data (i.e. *process* rather than *state* quality), or even with hardware. Participant examples of process

quality affecting perceived data quality were discussed in the previous section (Section 4.1.3), including processes such as data generation, delivery (e.g. interface), derivation, or use. A consultant in the practitioner focus group noted that many of the pragmatic criteria "such as *accessible* and *flexible* are heavily dependent on the underlying data model". An academic specializing in data management for decision support commented that "*accessibility"* may depend on different levels of technology, from the end-user interface in application software to system software or hardware such as database access structures—data indexes." Another practitioner noted the difficulty in distinguishing between data quality and other types of quality problems even in the case of objective quality assessment, commenting that "data and process quality are intertwined…data quality problems can be symptoms of process problems". A senior data quality manager for a large organization gave the example of "an insurance company [that] had an incorrect data structure whereby deceased people were not recognized and twenty widows received letters addressed 'Dear Deceased'…'we are soliciting more business'".

The current research is specifically intended to serve as a basis for developing quality assessments rather than problem source analysis; however, such assessments would be expected to facilitate such analysis both by identifying specific areas of concern (i.e. in terms of individual criterion ratings) and through comparison of objective and subjective quality judgements.

4.1.5 Unified vs. Non-Unified Views

Participants discussed commonly experienced problems related to non-unified views of data, i.e. when multiple disparate schematic views of data with respect to data definition, integrity constraints, documentation, or other meta-data co-exist in a data collection. One academic participant raised the case of the "same people or stakeholders dealing with different sets of data based on different standards or people moving between jurisdictions."

Disparate views of data can arise as a result of differing temporal, jurisdictional, contextual (e.g. geographical or administrative) perspectives. An academic specializing in conceptual data modelling noted that the "applicable legal acts constraining data change over time" and that the database may include "incompatible rules applicable to existing but not new data". He gave an example of "an extant insurance policy [that] depends on the rules valid at the time the policy was acquired even if [the rules are] no longer relevant to new policies, thus requiring incompatible rules be maintained in the database". A report and application developer in the practitioner focus group stressed that "organizations and organizational structure change over time" and questioned "what happens to the quality of the data when the data structure changes and the data warehouse is not updated accordingly?" In the individual opinion form, the same participant noted that conformance to database integrity rules "may not be possible if rules change but historical data is still present". Similarly, the validity of a specific schematic element (e.g. data definition, field) may have jurisdictional limitations. An academic participant noted that there are "different national and jurisdictional standards for determining gender…four in Australia and three in the USA". An academic in data management asked about "different perspectives within companies" affecting views of data and their definitions. As a result, the context of the data must be known and understood to determine the relevant descriptive (e.g. data definition) or constraining metadata (e.g. integrity rule). This complicates the assessment of data quality relating to conformance to metadata (i.e. objective and subjective assessments of syntactic data quality), since the value of any such judgements necessarily depends

on the existence of appropriate metadata and the quality of that metadata (i.e. data model quality). Such considerations are particularly significant when data has been removed from its original context (i.e. the context is not well-understood or known) or collated from multiple sources and temporal reference points, e.g. in data warehouses.

In this research, we assume that disparate data views either (1) have been effectively reconciled in a single unified view or (2) are managed transparently via an underlying support mechanism (e.g. by software supporting heterogeneous, federated, or warehoused data). In fact, this is just a further variation on the earlier assumption of perfect metadata in Section 3.1.2, justified by our focus on information rather than data model quality. To the degree that these assumptions do not hold, IQ will inevitably be impacted. However, as discussed in the previous section, quality assessments based on the proposed framework would be expected to facilitate discovery of such problems.

4.2 MISSING CRITERIA

Participants suggested a number of potential additions to the list of quality criteria. The data warehouse report and application developer from the practitioner focus group was particularly concerned with metadata quality, asking “how should we manage changes to the data model and the effects on data instances?” in the focus group and noting problems with “undocumented, out-of-date schemas” on the individual opinion form. An academic participant specializing in data mining suggested in her individual opinion form and in focus group discussion that the “level [comprehensiveness] of documentation” be additionally considered. Another academic participant specialized in decision support agreed that she was “also interested in the quality of metadata”. As discussed previously in Section 1, poor metadata quality can negatively impact information quality (i.e. be a source of poor information quality); however, the two terms are not synonymous. The proposed criteria *metadata quality* is a distinct topic (commonly referred to as *data model quality*) requiring separate analysis and treatment and thus considered to be outside the scope of the framework.

Two academic participants on the individual opinion form and two practitioners in the focus group discussion suggested that *cost* or *relative business impact* be included. A consultant thought that quality assessment should “highlight the expected value of loss related to the lack of data quality”. We regard this as a separate issue in itself, i.e. determining the business impact of quality problems or potential improvements. It is more closely related to the use of the framework and any assessment tools or techniques based on it (e.g. which data sets are assessed and how often) rather than to the appropriate content of the framework itself. For example, a senior data quality manager stressed the importance of “identifying critical set of business data” and “baselining” and periodically assessing their quality.

Since specialized data are judged to be out of the scope of the current research (see Section 4.1.1), the same is true of their specific criteria such as *representative* or *probable* (see Section 4.1.1).

On consideration, another proposed addition relating to *privacy* (by one academic participant in the individual opinion form) led to the elaboration of the definition of the pragmatic criterion *security*. The original definition of “appropriately protected from damage or abuse (including unauthorized access)” was amended to include unauthorized use or distribution as well as access. For example, the senior data quality manager from the large organization commented in the focus group

discussion on common privacy violations related to unauthorized distribution such as “giving information to the spouse” and further noted that “national privacy rules are contradictory”. We felt that any attempt to address the full legislative significance of *privacy* was outside the scope of the framework since it varies considerably between countries and with respect to the internal consistency and coherency (and relevancy to information quality) of the issues so labelled.

Only one proposed addition was both within the framework scope and not currently addressed by any criterion: *including access to appropriate metadata*. The same academic participant who raised the question of metadata quality further asked in the focus group discussion “whether users’ access to metadata is relevant to quality?” and, later, “is availability of metadata indicative of data quality?”. As it is clearly use-related, this criterion is added to the pragmatic category. In fact, this suggestion to include this quality criterion arose more than once in separate focus group feedback and was motivated by the requirements of different application contexts, including the following examples of requirements:

- for version documentation of replicate data, with the data warehouse report and application developer from the practitioner focus group noting on the individual opinion form that it is important to be able to identify “what version of the data we are working with” and commenting again in discussion that it is “critical to know what generation of copies a particular data item is”.

- for currency, lineage, granularity, transformation, and source documentation of spatial data, with the academic specialized in spatial data noting in the individual opinion form that such data is “largely heterogeneous…from different sources, have different histories, different quality” and that critical spatial data quality information “relating to lineage, currency, positional and attribute accuracy, logical consistency,… completeness…known error rates and details of processes used to assess these rates…are typically embedded in metadata”,

- for documentation on data collection purposes to comply with privacy legislation, with the comment in the academic focus group discussion that “the original purposes of collecting data must be known/retrievable to ensure that data is not used for different purposes as required by national privacy legislation”, and,

- for documentation of context for data originating from disparate or unfamiliar sources, especially in a data warehouse or data collection external to the organization accessing the data. The data warehouse report and application developer gave the example of “metadata information [that is] required in order to know which of three different ‘approved’ methods of calculating insurance premiums should be used”.

A data set that does not provide access to relevant metadata may result in data being unintelligible, misinterpreted, or unintentionally misused. This clearly impacts the perceived quality of the retrieved information. The first possible consequence listed, unintelligible data, further implies an inter-dependency which should be acknowledged between this new quality criterion and the quality criterion *understandable*.

Finally, it should be noted that academic participants were explicitly asked in the course of the focus group discussion "whether there was a need to include terms such as *believability, trust,* or *reputation* in the framework". The participants felt that (i) these terms were ambiguous and subject to differing interpretation and (ii) quality concepts implied by these terms were subsumed by current framework criteria.

4.3 INTER-DEPENDENCIES BETWEEN CRITERIA

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.

4.3.1 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 Section 3.2.2, the original semantic definitions, expressed in terms of states, were operationalized in terms of identifiable data units (i.e. IS data artefacts) and external phenomena. As a result, *correct* was initially defined as having attribute values match property values for each represented external phenomenon. However, an academic participant specializing in decision support identified in discussion an "overlap between *correct* and, if there is an incorrect identifier [key value] resulting in *ambiguous, redundant,* or *meaningless* mappings, other semantic criteria" and gave the example of a "name spelled incorrectly and correctly in the database as two separate instances for one person" resulting in both incorrect and redundant mappings. This relates to a consultant's comment in the practitioner discussion that "identity matching is required before checking detailed correspondence".

To eliminate this source of inter-dependency, two separate semantic correctness criteria, *phenomenon-correct* and *property-correct,* were defined. The first relates to the correctness of mapping identifiable IS data units to external phenomena based on keys (i.e. identification information). A violation would involve an unambiguous, meaningful, non-redundant mapping of an identifiable data unit to the *wrong* external phenomenon. The second 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 phenomena-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.

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 property value. For example, a consultant in the practitioner group gave the example of "an address that violates Australia Post Standards but can be correctly interpreted by call center personnel". We therefore clarify that *property-correct* is with respect to fidelity to external property values, but not necessarily to all specified integrity rules.

4.3.2 Pragmatic Criteria

In the individual opinion form responses, evaluation-dependencies were identified between pragmatic criteria relating to data delivery, in that information must first be *accessible* to judge whether it is *understandable* (two academic participants) and must be *understandable* before judging whether it is *suitably* (three practitioner and four academic participants) and *flexibly presented* (two practitioner and two academic participants). Logically the converse dependencies, which are value-dependencies, must also be considered; namely, information presentation could potentially affect perceived understandability. In contrast, there was complete consensus in individual opinion forms that *accessible* was independent of other criteria. 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-criterion *timely* was removed from *suitably presented* and made a separate delivery-related criterion. This restricts the criterion *suitably presented* to presentation style (i.e. layout, precision, units) aspects, thus simplifying and clarifying its semantics. Furthermore, it serves to acknowledge the critical importance of timeliness as a quality aspect in its own right. This was an issue raised in both focus groups, independently by two separate academics and by one practitioner in individual opinion forms and by the same academics in discussion. For example, an academic specializing in decision support commented verbally that “timeliness is not a sub-aspect, it is important enough to be a separate criterion”.

Further evaluation-dependencies between *understandable* and many other criteria (beyond those relating to data delivery) were identified. Essentially, as noted by a consultant in the individual opinion form, 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 not covered elsewhere in the framework. This was explained in Section 3.2.3 and illustrated there using the example of *lineage* as a quality criterion specific to spatial applications. However, this represents a specific example of the more general quality criterion added to the framework as a result of focus group feedback, *including access to relevant metadata* (discussed in Section 4.2). Focus group discussion failed to elicit any examples of such domain-specific criteria that did not fit into the modified framework, 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. In individual opinion forms, three practitioners and five academics noted inter-dependencies between *relevant* and *valuable*. An academic specializing in decision support commented on the form that “*valuable* is virtually meaningless—too high a level of abstraction representing a combination of other concepts, e.g. *relevant, usefulness, sufficiency, timeliness*” and in discussion that “*valuable* is not useful because it is too abstract”. Another academic specializing in spatial data noted in the individual opinion form that “you cannot separate different influences of

each user's concept of *valuable"*. A data quality manager commented in the individual opinion form that "I don't understand the distinction between *relevant* and *valuable*" and in the practitioner focus group discussion that he "got confused with the difference between *relevant* and *valuable*". 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 and therefore not useful as a specific quality criteria. Therefore, it was decided that the criterion *valuable* should be removed.

It was additionally observed by the academic specializing in spatial data that although the framework did encompass spatial quality criteria (e.g. *lineage* as a specific example of *including access to metadata*), 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.

In a related issue, focus group feedback highlighted the fact that the *sufficiency* aspect, included in the criterion *valuable*, should, in fact, be considered in the criterion *relevant* with respect to the types of information available. Two academics noted in their individual opinion forms that *sufficiency* was the only important and distinct aspect of the criterion *valuable* and that it was closely related to *relevant*. (This is a plausible explanation of why eight out of fourteen respondents in the individual opinion forms rated the importance of *valuable* as *high*.) This aspect of quality is not considered elsewhere in the framework, since *relevant* considers only whether the data types available are pertinent rather than sufficient for the intended information use. Therefore, the criterion *relevant* is 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 (i.e. work task).

4.4 CRITERIA DEFINITION

In this section, we discuss focus group feedback relating to identified ambiguities in criteria meaning or wording not caused by dependencies between criteria. Identified ambiguities in wording were resolved by a change in terminology either for the nomenclature or for the definition of the criteria in question. Ambiguities in semantics were resolved by modifying the semantics and possibly the terminology of the criteria concerned. The majority of academics and practitioners indicated in the individual opinion forms that the criteria were clearly defined and important (i.e. ticked the *yes* and *high* boxes respectively), with the exception of the criterion *non-redundant* with respect to importance and the criterion *meaningful* with respect to clarity. These exceptions are discussed in Sections 4.4.1 and 4.4.2 respectively.

4.4.1 Ambiguities in Criteria Meaning

With respect to criteria meaning, it was evident from both 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 (i.e. controlled rather than uncontrolled redundancy), as an integral part of effective organizational data management. An academic with decision support specialization commented in the individual opinion form that "non-redundancy is not important for end-users—sometimes redundancy makes a system easier to use, e.g. in a data warehouse". A consultant noted in the practitioner focus group discussion that "redundancy may be desirable, so *non-redundancy* may not be quality criterion".

The argument presented in Section 3.2.2 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 re-define the quality criterion *non-redundant* as *consistent*, i.e. 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. Thus the replication specific issues of replicate consistency and update lag time are specifically addressed in the basic definition of the quality criterion *consistent*.

With respect to combining the semantic criteria *correct, meaningful, unambiguous,* and *non-redundant* into a single summary term *reliable*; four academics and two practitioners disapproved because valuable information would be lost and/or the summary term *reliable* would be less consistently comprehensible. In the individual opinion form, an academic specializing in data mining explained that "separation facilitates better explication and clarification of concepts", another academic specializing in decision support noted that "each criterion addresses a different aspect of data reliability…[having] different importance (weight) in the context of different applications, and a consultant practitioner felt that "by combining the items it may be more difficult to identify why an item may not be considered reliable". An academic specializing in spatial data noted that "you cannot interpret the user's notion of *reliability*—which is a vague term" and a senior data quality manager that "*reliability* has too many interpretations".

In contrast, one academic and two practitioners indicated in their individual opinion forms that combining semantic criteria would not necessarily be a problem. A consultant commented that "it depends on the context, I see no problem with using a summary term in some circumstances", an academic specializing in decision support wrote that it "depends on your purpose", and the data warehouse application and report developer felt that "the change would have little or no impact on the validity or useability of the methodology [framework]". Furthermore, an examination of the explanatory comments for and perceived inter-dependencies between the four semantic criteria in individual opinion form indicates that it was quite difficult for respondents to distinguish between the different criteria and mapping cardinalities defined. As a result, participants often misunderstood the criteria definitions. For example, the majority of respondents felt that there were inter-dependencies between the criteria *unambiguous*, *meaningful,* and *non-redundant* whereas if their different cardinalities were understood correctly then it would be logically clear that they were independent. Written explanations further demonstrated the level of confusion, with respondents querying whether "ambiguous data suggestive that data is incomplete?" and whether *non-redundant* "sort of

overlaps with *complete*" and commenting that they "didn't understand the question". The challenge is then how best to address the contradictory concerns of information loss versus confusion. On consideration, the decision was to retain the four separate criteria in the semantic category of the framework to provide detailed information (where they are objectively assessed) while combining them in the pragmatic category to avoid respondent confusion (where they are subjectively assessed through consumer perceptions). The new pragmatic criterion *reliable* is defined as "each identifiable data unit maps correctly to exactly one external phenomenon and each external phenomenon is represented consistently in the database (i.e. either by only one identifiable data unit or by multiple acceptably consistent identifiable data units)."

4.4.2 Ambiguities in Criteria Wording

Other identified ambiguities related to criteria nomenclature or the wording used in definitions. For example, many participants thought that the criterion *meaningful* related to the data's business importance (i.e. significance); whereas the definition given was related to the data's external mapping (i.e. mapping cardinality). Thus the implicit connotations of the English word took precedence over the explicit definition. Therefore the names of all the semantic criteria were amended to include explicit references to mapping, e.g. *mapped meaningfully, mapped completely*, etc.

4.5 SUBJECTIVE QUALITY ASSESSMENT TECHNIQUES

These issues relate generally to subjective IQ assessment based on the criteria defined for the pragmatic category of the framework. Participant concerns can be categorized as relating either to the understanding and evaluation of pragmatic framework criteria for a given data set or to the analysis of responses based on such criteria. We suggest possible ways to address each concern.

With respect to evaluation concerns, some participants felt that consumers might have difficulty evaluating subjective quality aspects relating to rule conformance (see Section 4.1.3), security, and replicate consistency (see Section 4.4.1)—since they might generally be unaware of these aspects except in cases of obvious breeches. For example, the conceptual model specialist commented in the academic focus group discussion "how can data consumers judge security? It may relate to processes using the database rather than the database itself". However, a majority of participants agreed (as evidenced by the individual opinion form importance ratings) that these were, nonetheless, important quality characteristics that needed to be included in the framework. This could be addressed by allowing *don't know* or *not applicable* responses in the evaluation of any such criteria, essentially permitting an opt-out for those individuals to whom the criteria in question do not seem relevant. Participants in the practitioner focus group further questioned whether consumers could understand the proposed criteria unless described in terms specific to each consumer's individual business responsibilities. A senior data quality manager "had trouble putting the individual opinion form into context and felt it was very technically oriented". Stressing the need to translate data quality terms and statistics into business issues, the manager gave the example of describing data relevance and completeness to telephone salespeople in terms of quicker, more effective sales calls. This is a valid concern that highlights the difficulty of developing an IQ framework of general utility across different application contexts without sacrificing context-specific understandability or relevance. One way to address this concern would be to include illustrative (or context-specific) examples with criteria definitions or by customizing the framework (including criteria

definitions) for specific application contexts. For example, a data quality consultant suggested that “each item [framework criterion] should be accompanied with an example”. Another concern is the possibility that sub-criteria within a single criterion might have different quality ratings, requiring separate assessment and appropriate techniques to collect and analyze subjective feedback [Straub et. al., 2004]. For example, with respect to the *accessible* sub-criteria *easy* and *quick,* we have the individual opinion form comment from a practitioner (responsible for application and report development) that “data can be easy to retrieve without being quick”.

With respect to questions regarding analysis of responses to framework criteria evaluation, participants raised concerns that averaging consumer responses might invalidate assessment significance. For example, an academic specializing in decision support and data warehousing commented in discussion that “averaging responses over a set of respondents may neutralize responses and render feedback meaningless” and gave the example that “managers think it is great and end-users think it is terrible, the average will be OK but that doesn’t give a good picture of bipolar role-dependent opinions”. It is our considered opinion that these concerns can generally be addressed through the flexible partitioning of responses (i.e. the individual partitions used to aggregate and average responses) during analysis of responses based on background information (e.g. with respect to organizational role, experience, position) provided by respondents. Further concerns related to whether such subjective assessments (1) take into consideration trade-offs between different quality criteria or (2) should be applied only to a set of critical business data (see also Section 4.1.3). An academic specializing in decision support and data mining noted in discussion that “criteria may involve trade-offs and contradictions” and that it is “important to identify critical business data”. We believe that these concerns can be addressed respectively based on the (1) presentation of results (i.e. by presenting results for individual quality criterion separately as appropriate) and the (2) administration of the subjective IQ assessment (i.e. by selecting an appropriate target data set for the assessment). In fact, these concerns and their responses are more generally relevant to any subjective IQ assessment based on a multi-criteria IQ framework, regardless of the specific framework used.

4.6 OBJECTIVE QUALITY ASSESSMENT

These issues relate generally to objective IQ assessment based on the criteria defined for the syntactic and semantic categories of the framework. The focus was on the practical difficulties of evaluating conformance and correspondence respectively for these two categories of criteria.

The difficulty of establishing the correct mapping of IS data artifacts to external phenomena and of confirming detail correctness was discussed at length in the first focus group. Based on extensive professional experience, participants highlighted the subjectivity (variability) of correspondence judgments and the difficulty in finding authoritative sources against which to compare data. A consultant stressed that “establishing database/real-world correspondence is problematic and context-dependent” and that “people may not agree on correspondence rules”. A senior data quality manager in a large organization asked “baseline your data against what?...there is no reference point in Australia against core data attributes” and that “there are very few authoritative sources. The manager described methods useful in checking data correctness including (1) “comparing agreement between multiple independent external sources or internal replicates” and (2) “source verification (either deliberate or piggybacked with other customer contacts such as

those made for sales or marketing purposes)". A consultant described similar problems in checking data conformance to rules, i.e. the difficulty in establishing an authoritative and commonly accepted set of rules for a data set. In fact, these concerns represent an integral aspect of any objective quality assessment regardless of the specific IQ framework used as a basis for the assessment.

5. REVISED FRAMEWORK

As a result of focus group feedback, the scope of the semiotic IQ framework discussed in this report can be clarified as follows. The framework is specifically intended for general business applications and for use with electronic data sets that include identifiers (i.e. key attributes) allowing mapping between IS data and external phenomena.

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 definitions for each quality category are shown in Table 3 below. Where relevant, sub-criteria are listed in parenthesis after the criterion name.

Table 3. Revised Quality Criteria by Category

Syntactic Criteria (based on rule conformance)
Conforming to integrity rules. Data follows specified database 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.
Mapped correctly to external phenomena. Each identifiable data unit maps to the correct external phenomenon.
Mapped correctly to external properties. 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 is easy and quick to retrieve.
Suitably presented (suitably formatted, precise, and measured in units). Data is presented in a manner appropriate for its use, with respect to format, precision, and units.
Flexibly presented (easily aggregated; format, precision, and units easily

converted). Data can be easily manipulated and the presentation customized as needed, with respect to aggregating data and changing the data format, precision, or units.
Timely. The currency (age) of the data is appropriate to its use.
Understandable. Data is presented in an intelligible manner.
Secure. Data is appropriately protected from damage or abuse (including unauthorized access, use, or distribution).
Type-sufficient. The data includes all of the types of information important for its use.
Includes access to appropriate metadata. Metadata is available as appropriate to define, constrain, and document data.
Conforms to integrity rules (consumer perception). Data follows specified database integrity rules
Mapped completely (consumer perception). Every external phenomenon is represented.
Mapped reliably (consumer perception). Each identifiable data unit maps correctly to exactly one external phenomenon and each external phenomenon is represented consistently in the database (i.e. either by exactly one identifiable data unit or by multiple acceptably consistent identifiable data units).

The revised quality criteria in Table 3 can be compared with the initial list of quality criteria in Table 2 to identify those theoretically-derived criteria most affected by the empirical refinement process. The syntactic category was not affected. In the semantic quality category, the most significant changes are:

- the division of the *correct* criterion into two separate criteria describing different aspects of correctness, and
- the re-definition of the *non-redundant* criterion in terms of consistency.

Similarly, we can compare the revised quality criteria in Table 3 to the initial list of quality criteria in Table 2 to identify the pragmatic criteria most significantly changed:

- the new treatment of *timely* as a separate criterion rather than a sub-criterion of the *suitably presented* criterion,
- the elimination of the *relevant* and *valuable* criteria and replacement with the criterion *type-sufficient*, and
- the addition of the criterion *including access to appropriate metadata*.

6. CONCLUSION

This report presents a comprehensive description of the development of an IQ framework intended to address concerns related to rigor without sacrificing those related to scope. To achieve this, we adopt an approach to developing a framework based on the use of semiotic theory as a theoretical foundation for (1) defining quality categories, (2) rationalizing the derivation of category criteria, and (3) classifying quality criteria. The first step provides a logical basis for defining and

differentiating between the semantics of different quality categories. Furthermore, the fact that the third step is an implicit (i.e. automatic) consequence of the first two steps ensures a consistent classification of criteria. To our knowledge, no other IQ research to date provides a theoretical basis for defining quality categories and classifying criteria into those categories. Semiotic theory further provides a basis for integrating objective, product-based and subjective, service-based quality perspectives in one coherent, unified framework. Thus, the use of semiotics addresses problems in other related work with respect to scope or inconsistency.

Empirical work based on focus groups served to refine the framework, especially the subjective quality criteria. Specifically, the feedback serves to clarify the scope (i.e. applicability) of the framework as a whole; to identify inter-dependencies, ambiguities, and omissions in the initial set of criteria; and to highlight potential benefits and implications of applying the framework to IQ assessment or management.

The explicit intention of this research was to provide an IQ framework that could serve as a basis for further work in IQ in general and in IQ 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 IQ such as improvement and management and to specialized application contexts involving, for example, spatial or scientific data.

REFERENCES

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

Bunge, M. (1977) *Treatise on Basic Philosophy: Vol 3: Ontology I: The Furniture of the World,* Boston: Reidel.

Bunge, M. (1979) *Treatise on Basic Philosophy: Vol 4: Ontology II: A World of Systems,* Boston: Reidel.

Chengular-Smith, I.N., Ballou, D. and Pazer, H.L. (1999) "The Impact of Data Quality Information on Decision Making: An Exploratory Analysis", *IEEE Transactions on Knowledge and Data Engineering*, (11)6, (November/December)

Chisolm, R. (1996) *A Realistic Theory of Categories--An Essay on Ontology*, Cambridge: Cambridge University Press.

Dyke, T. P. V., L. A. Kappelman, and V. R. Prybutok (1997), "Measuring Information Systems Service Quality: Concerns on the Use of the SERVQUAL Questionnaire", *Management Information Systems Quarterly* (21)2, pp. 195-208.

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), pp. 167-182.

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

Kahn, B. K., D. M. Strong, and R. Y. Wang (1997) "A Model for Delivering Quality Information as Product and Service", *Proceedings of Conference on Information Quality*, Cambridge; Massachusetts Institute of Technology, pp. 80-94.

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

Krogstie, J. (2001) "A Semiotic Approach to Quality in Requirements Specifications", *Proceedings of IFIP 8.1 Working Conference on Organizational Semiotics*, Montreal, Canada, pp. 231-249.

Krogstie, J., O. I. Lindland, and G. Sindre (1995) "Defining Quality Aspects for Conceptual Models", *Proceedings of IFIP8.1 working conference on Information Systems Concepts (ISCO3): Towards a Consolidation of Views,* Marburg, Germany, pp. 216-231.

Krueger, R.A. (1994) *Focus Groups: A Practical Guide for Research*, Thousand Oaks, CA: Sage.

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

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

Parasuraman, A., L. L. Berry, and V. A. Zeithaml (1991), "Understanding Customer Expectations of Service", *Sloan Management Review* (32)3, pp. 39-48.

Parasuraman, A., V. A. Zeithami, and L. L. Berry (1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality", *Journal of Retailing,* (64)1, pp. 12-40.

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

Pitt, L. F., R. T. Watson, and C. B. Kavan (1997), "Measuring Information Systems Service Quality: Concerns for a Complete Canvas", *Management Information Systems Quarterly*, pp. 209-221.

Shanks, G. and P. Darke (1998a) "Understanding Data Quality in Data Warehousing: A Semiotic Approach", *MIT Conference on Information Quality,* Boston, pp. 247-264.

Shanks, G. and P. Darke (1998b) "Understanding Metadata and Data Quality in a Data Warehouse, Australian Computer Journal", (30)4, pp. 122-128

Straub, D., M.C. Boudreau, and D. Gefen (2004) "Validation Guidelines of IS Positivist Research", *Communications of the ACM* (14), pp. 380-426.

Stamper, R. (1991) "The Semiotic Framework for Information Systems Research"' *Information Systems Research: Contemporary Approaches and Emergent Traditions*, Amsterdam: North-Holland.

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

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

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

APPENDIX 1

Individual Opinion Form for Data Quality Focus Group Participants:

Welcome to our focus group and thank you for your time. Please allow ¾ hour to complete this form. The information you provide will help us prepare for the focus group.

INSTRUCTIONS

1. Please **read the Background information section** carefully. This describes the Focus Group Questionnaire context by giving a brief overview of the information quality framework of which it forms a part.
2. Please **read the entire Focus Group Questionnaire section before answering the questions**, as some of the questions require familiarity with the entire Focus Group Questionnaire.
3. After reading the Focus Group Questionnaire, please **answer all 15 questions**:
   a. Questions 1-13 each contain underlined sub-questions asking whether the listed criterion is *clearly defined*, *independent* (except see Note2), and *important.*
      i. **Please check only ONE box for each underlined sub-question.**
      ii. **If you answered** *No* **or** *Low* **for any sub-question, please give specific reasons for that answer in the** *Comments* **section.** If you need extra space for comments, either insert extra lines (softcopy) or add to the end (hardcopy), labeling extra comments by the question number to which they refer.
4. We would very much appreciate if you could **fax the completed Focus Group Questionnaire to us before the focus group discussion.**

BACKGROUND INFORMATION

Continuous measurement, improvement, and management of data quality[5] are essential to the success of an organization; however, we have not yet fully answered the pre-requisite question of how to define data quality in an organization. A widely accepted high-level definition of quality is "fitness for use or purpose". However, a more detailed definition giving specific criteria (i.e. quality characteristics) is required for practical purposes such as data quality assessment. Current proposals in

[5] Information or data quality commonly refers to the quality of information system data (the topic of this focus group) rather than the quality of the data model per se. Although data quality cannot be completely separated from data model quality (e.g. some data quality judgements relate to type-based, vertical, partitions of the instance data); in general, these represent two distinctly different concepts requiring separate analysis and treatment.

the literature define quality in terms of individual criteria subsequently grouped into general quality categories. Empirical, ad-hoc, or theoretical research approaches have been used to derive individual quality criteria and their groupings into categories. The first two approaches have typically been used as a basis for assessing perception-based aspects of quality related to data use; however, they have been criticised as lacking rigor (i.e. having internal contradictions or omissions), especially with respect to the selection and definition of general quality categories. In contrast, theoretical approaches proposed to date have better consistency and rigor but are limited in scope to the more objective and static aspects of quality such as integrity rule compliance or real-world correspondence.

To address these concerns, we have proposed that an initial focus on the theoretical derivation of general quality categories can improve overall rigor and facilitate integration of subjective and objective quality aspects. In particular, we have proposed an information quality framework and theoretically-based quality category derivation based on concepts from semiotic theory—the theory of signs. A datum in a data collection can be regarded as a *sign* or representation of some component of the real-world e.g. an employee record as a representation for a company employee. Following from this observation, 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 real-world, and (3) suitability for use. Continuing the earlier example, 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 correctly represents an real employee (e.g. has matching details), and (3) available employee information is useful for the tasks performed by the information consumer accessing the data collection.

Whereas the syntactic and semantic quality categories are amenable to a theoretical analysis (not described here) and can be measured by fairly objective means (e.g. analysis of the data collection and sampling respectively); the pragmatic quality category requires at least a partly empirical approach and its assessment involves subjective measurements of consumer perceptions. **The Focus Group Questionnaire below relates to the development of a general survey instrument**

(i.e. a consumer questionnaire) that can be used to assess pragmatic quality based on consumer judgement.

From a theoretical and literature-based analysis of data quality, we have developed an initial list of pragmatic data quality criteria intended for use in developing a consumer questionnaire. The intended context of the Focus Group Questionnaire is for a single consumer, task, and data set (i.e. the data set used to complete that task by that consumer). Note that the pragmatic quality assessment includes a component involving subjective, perception-based assessment of the same criteria that are measured objectively at the syntactic and semantic levels. That is, the first six pragmatic criteria listed in the Focus Group Questionnaire below correspond to the theoretically-derived syntactic (criterion 1) and semantic (criterion 2-6) level criteria; however, they will be used to solicit subjective consumer judgements rather than used as the basis of objective measurement. This allows direct comparison of objective and subjective quality assessments.

The Focus Group Questionnaire below solicits your opinion as to the efficacy of the initial list of pragmatic criteria proposed. Your feedback will help us to identify any unclear, invalid, overlapping or missing data quality criteria. In particular, you will be asked to evaluate a list of data quality criteria and their definitions with respect to whether they are (1) **clear**, (2) **independent**, (3) **valid**, and (4) **comprehensive**, described as follows:

- *Clear* means that the definition of each criterion is easily understandable, unambiguous, and matches your intuitive understanding of the name used for that criterion.
- *Valid* (i.e. important) means that every criterion on the list is a good description of some significant aspect of data quality. In the Focus Group Questionnaire, we use the term *important* to refer specifically to the significance of the described criteria.
- *Independent* (i.e. non-overlapping) means that, as far as possible, each criterion should describe a unique aspect of quality not described by any other criteria in the list. In other words, each significant aspect of data quality is described by only one criterion in the list.
- *Comprehensive* means that all aspects of data quality have been included in the list, i.e. that there are no significant data quality aspects that are not described by or subsumed by (i.e. included in) at least one of the criteria in the list.

FOCUS GROUP QUESTIONNAIRE

For each data quality criterion listed below, please **check only ONE box** per underlined sub-question and, if you answer *No* or *Low*, **explain why** in the *Comments* section.

1. *Conforms to Rules:* data should follow specified database integrity rules.

Clearly defined?	**Independent?**	**Importance?**
Yes ☐ No ☐	Yes ☐ No ☐	High ☐ Medium ☐ Low ☐

Comments:

Note1: Criteria 2 through 6 can be summarized by the term *Reliable*, i.e. there is a one-to-one correspondence (i.e. match) between database information and represented **real-world instances (i.e. phenomenon, e.g. an object such as a specific employee or property value such as age 29)**. (In other words, the cardinality of mapping data to real-world instances must be 1-1 with database attribute values matching real-world property values).

2. *Complete:* every real-world instance relevant to your work is represented in the database. (In other words, the cardinality of mapping data to real-world instances cannot be *zero* on the data side, i.e. can only be 1 or M data to 0, 1, or M real-world instances.)

Clearly defined?	**Independent?**	**Importance?**
Yes ☐ No ☐	Yes ☐ No ☐	High ☐ Medium ☐ Low ☐

Comments:

3. *Correct:* database details (i.e. attribute values) should exactly match details (i.e. property values) of the corresponding (i.e. represented) real-world instance.

Clearly defined?	**Independent?**	**Importance?**
Yes ☐ No ☐	Yes ☐ No ☐	High ☐ Medium ☐ Low ☐

Comments:

4. *Unambiguous:* each datum or set of datum in the database represents at most one specific (i.e. a single) real-world instance relevant to your work. (In other words, the cardinality of mapping data to real-world instances cannot be *many* on the real-world side, i.e. can only be 0,1, or M data to 0 or 1 real-world instance.)

Clearly defined? Yes ☐ No ☐ **Independent?** Yes ☐ No ☐ **Importance?** High ☐ Medium ☐ Low ☐

Comments:

5. *Meaningful:* each datum or set of datum in the database represents at least one specific (i.e. a single) real-world instance relevant to your work. (In other words, the mapping of data to real-world instances cannot be *zero* on the real-world side, i.e. can only be 0, 1, or M data to 1 or M real-world instances).

Clearly defined? Yes ☐ No ☐ **Independent?** Yes ☐ No ☐ **Importance?** High ☐ Medium ☐ Low ☐

Comments:

6. *Non-Redundant:* each real-world instance relevant to your work is represented only once in the database (In other words, the mapping of data to real-world instances cannot be *many* on the data side, i.e. can only be 0 or 1 data to 0, 1, or M real-world instances.)

Clearly defined? Yes ☐ No ☐ **Independent?** Yes ☐ No ☐ **Importance?** High ☐ Medium ☐ Low ☐

Comments:

7. *Relevant:* the types of data available are pertinent to your work tasks.

Clearly defined? Yes ☐ No ☐ **Independent?** Yes ☐ No ☐ **Importance?** High ☐ Medium ☐ Low ☐

Comments:

Note2: Criterion 8 below is somewhat dependent on all the other criteria, since if data does not satisfy other criteria then the data will necessarily be somewhat less valuable. However, even if all the other criteria are satisfied; the data may still not be valuable for a given work-task if criteria specific to the business domain are not satisfied. Essentially, criterion 8 acts as a generic placeholder for business-specific quality criteria.

8. *Valuable:* the types of data available are useful (i.e. important) and sufficient for your work tasks.

Clearly defined? **Independent?** **Importance?**

Yes ☐ No ☐ Yes ☐ No ☐ High ☐ Medium ☐ Low ☐

Comments:

Note3: Criteria 9 through 13 refer to the presentation and delivery of data.

9. *Understandable:* data is presented in a manner easy to interpret.

Clearly defined? **Independent?** **Importance?**

Yes ☐ No ☐ Yes ☐ No ☐ High ☐ Medium ☐ Low ☐

Comments:

10. *Accessible:* data is easy and quick to retrieve.

Clearly defined? **Independent?** **Importance?**

Yes ☐ No ☐ Yes ☐ No ☐ High ☐ Medium ☐ Low ☐

Comments:

11. *Secure:* data is appropriately protected from damage or abuse (including unauthorized access).

Clearly defined? **Independent?** **Importance?**

Yes ☐ No ☐ Yes ☐ No ☐ High ☐ Medium ☐ Low ☐

Comments:

12. *Flexible:* data can be easily manipulated and the data presentation customized as needed, including aggregation and format (e.g. unit, precision, representation) conversions.

Clearly defined? **Independent?** **Importance?**

Yes ☐ No ☐ Yes ☐ No ☐ High ☐ Medium ☐ Low ☐

Comments:

13. *Suitable:* the data is presented in a manner suitable for your work tasks; including timeliness, quantity (the amount of data), format, precision, units.

Clearly defined?	**Independent?**	**Importance?**
Yes ☐ No ☐	Yes ☐ No ☐	High ☐ Medium ☐ Low ☐

Comments:

14. Are all significant aspects of data quality included explicitly in the list? If not, **please specify which criteria are missing and define them in your own words.** This includes any aspect of data quality which is subsumed by (i.e. included in) one of the 13 explicitly listed criteria (**please specify which)**, but you think it should be listed explicitly as a separate criterion.

15. Do you think that it would be more effective to replace criteria 3-6 (i.e. *correct, meaningful, unambiguous, non-redundant*) with the single summary data quality criterion *reliable*? **Please explain why or why not?**

APPENDIX 2

Suggested Data Quality Focus Group Discussion Plan

1) DQ Focus Group Plan

a) Brief discussion of focus group & DQ project
   i) participants anonymous in any reporting of focus group results
   ii) participants will receive summary of focus group findings
   iii) data stored securely for 5 years and then destroyed (Monash ethics compliant)

b) Introduce focus group discussion plan:
   i) introduction,
   ii) core topics (item 3 below, i.e. defining DQ)
   iii) topics driven by participant interests (4-6 below or additional as suggested)

c) Any questions from participants on any aspect of focus group?

2) Introductions /Background for each person

a) For our "team" members (name, position/background, role in project & focus group)

b) For focus group participants (name):
   i) job function
   ii) experience with data quality
   iii) their most important data quality issue

3) Defining Data Quality

a) Clarify that we want general DQ criteria, i.e. valid across applications, to allow:
   i) pre-specification of generally applicable DQ measurement tools/techniques (e.g. DQ evaluation questionnaires)
   ii) possible DQ benchmarking across organizations

b) For each criterion: definition or name unclear (or unintuitive)?
   i) Why?
   ii) How improve?

c) For each criterion: invalid (i.e. either not a data quality concern, of very low importance, or not a correct description of the data quality characteristic)?
   i) Why?
   ii) How improve?

d) Any unidentified dependencies between listed criteria (other than the identified dependency of valuable on all the other criteria)?

e) Any missing criteria?

f) Should any of the "sub-items" under *flexible* and *suitable* criteria be considered criteria in their own right? Why or why not?

g) Should *reliable* be used instead of criteria 3-6? Why? Why not?

h) Discussion of *valuable* (#8)
   i) Are all DQ criteria general or can they think of some DQ criteria specific to data for a particular business or domain? (Can we think of an example?)

ii) Do they agree with my view that *valuable* serves as a "catch-all" for any such domain-specific DQ criteria? That is, we believe that even if data satisfies all the other listed DQ criteria, it may still be lacking in quality because specific DQ criteria are not satisfied (i.e. hence the data is not "valuable"). Is there another solution to including *valuable* other than specifying specific criteria for each business domain?
iii) Could we omit *valuable* and still be comprehensively defining DQ? That is, we believe that if data meets all the other listed criteria it must necessarily be of high quality.

Note: Further discussion issues will be selected from those below or added as suggested based on participant interests.

4) Measuring Data Quality in Practice

a) Perception-based versus objective sampling-based Methods of Measuring DQ
   i) Which would they trust?
   ii) Which would be most useful for justifying expenditure?
   iii) What authoritative source could be used to establish relevant real-world populations for sampling?

b) For accurate DQ assessment/measurement, is it important to be able to specify the relative importance of different DQ criteria (e.g. give criteria different weights or priorities) for a given business task and/or data set?

c) Which methods (if any) are used in their organization? Are they effective? Are they well-received? What justification was used?

d) Are any tools used to measure DQ?

5) Identifying, Determining sources of, Determining costs of, and Solving DQ Problems in Practice

a) How is this done in their organization? Is it: effective? well-received? justified?
b) How do they think this should be done ideally?
c) How important are these tasks for justifying expenditure?

6) Managing Data Quality in Practice

a) How is this done in their organization? Is it: effective? well-received? justified?
   i) Is there a separate DQ unit or is it part of another business unit (if so, which)?
   ii) What is the size & budget of the DQ unit?
   iii) Who does the DQ unit report to?
   iv) How is DQ management scoped/prioritized (e.g. by market segment, critical business function, CEO whim)?
   v) Are any specific tools used to manage DQ?
   vi) How is expenditure on DQ justified?
   vii) Is DQ management effective and is it well-received?