BUSINESS PROCESS MODELING TOWARDS DATA QUALITY ASSURANCE

An Organizational Engineering Approach

Keywords: Data Quality, Business Processes, Object-Oriented Modeling, UML

Abstract: Data is produced and consumed everyday by information systems, and its inherent quality is a fundamental aspect to operational and support business activities. However, inadequate data quality often causes severe economic and social losses in the organizational context. The problem addressed in this paper is how to assure data quality, both syntactically and semantically, at information entity level. An information entity is a model representation of a real world business entity. To address this problem, we have taken an organizational engineering approach, consisting in using a business process-modeling pattern for describing, at a high level of abstraction, how to ensure and validate business object data. The pattern defines a conceptual data quality model with specific quality attributes. We use object-oriented concepts to take advantage of concepts such as inheritance and traceability. The concepts and notation we use are an extension to the Unified Modeling Language. A case study is detailed exemplifying the use of the proposed concepts.

1. INTRODUCTION

Assuring data quality is a complex process, in which the tradeoff between cost and quality depends on the application context and on the requirements of an organization.

Incorrect, incomplete or non-timely data may cause social and economic problems in organizations, which very often, only react to its consequences, rather then having a proactive attitude.

Another common issue is that data quality is not understood in a process-centric, cross-departmental perspective, but at best in a functional view, as a duty or competence of an organization information systems department.

An organization may follow strict guidelines in the adoption of data quality projects, identifying critical problems and developing business processes and metrics for auditing or continuous improvement. There are several pragmatic and very successful approaches, but they are either focused at low level data analysis, based on computing algorithms treating implicitly data quality at the DBMS level, or they are mainly focused on quality management systems, based on ISO standards. These approaches are however not sufficient from the data consumer's point of view, (Laudon, 1986, Wang, 1993a).

Problems with data quality occur widely on functional organizations, where specific databases are created, forming information islands that constitute one of the mainstream problems, causing lack of consistency and of coherence of corporate data. Data might lack veracity in content and semantics. That is, data is not always correct syntactically and semantically, as the organization information system requires.

We propose a different approach from the one usually taken in the database research field, by applying business processes modeling and multi-contextual data quality concepts to actively pursue data quality objectives, within the information system of an organization.

The remaining of this paper is structured as follows: in section 2, we present the data quality concepts needed to attain our proposal. In section 3, we define the problem and propose our approach to its resolution. In section 4, we illustrate an example

of business process and data quality modeling and finally in section 5 we present some conclusions.

2. DATA QUALITY APPROACH

Data are generally facts about real world events, representing things and other business entities by means of symbols or other kind of representations.

Data might also be attributes, being the compound of names, means, valid values and business rules that guide integrity and correction.

Data may be considered as raw material for information product manufacturing, in analogy with the physical product production process of any industry, (Wang, 2001). Just as is difficult to manage product quality without understanding which product attributes define its quality, it is equally difficult to manage data quality without understanding what characteristics define it. Is thereafter necessary to understand what data quality is.

Data quality is associated with the data itself much in the same way as the quality of a product is associated in the mind of the consumer with the product itself.

2.1 Data Quality Requirements

Data quality requirements are distinct from application requirements and from application development quality requirements. Data quality requirements are critical in legacy systems, since often they were not designed to assure data quality In newer systems, there is also lots of subjectivity in deciding which data quality dimensions to include and their respective attributes.

In the past, data quality was often defined as non-conformance to requirements, (Crosby, 1984). However, just as there are several dimensions of product quality, such as conformance, durability or performance, in any industrial domain, data quality embraces specific characteristics, denominated quality dimensions.

The key dimensions of data quality (Table 1) are accuracy, completeness, consistency, currency and timeliness (Huh 1990, Ballou 1987).

2.2 Data Quality Modeling

Data quality modeling is somewhat like data modeling but focused on quality aspects, from where can be an analogy on requirements analysis methodology or database lifecycle, (Wang, 1993b).

Data quality is a multi-dimensional and hierarchical concept, (Wang, 1995). Data from several sources may have common dimensions in which quality may be measured. There are four categories and sixteen identified dimensions in the fitness for use by the information consumers, (Huang, 1998).

Category	Data Quality Dimensions
Intrinsic	Objectivity, Believability, Reputation, Accuracy
Accessibility	Access , Security
Contextual	Relevancy, Added Value, Timeliness, Completeness, Amount of Information
Representational	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation, Ease of Manipulation

Table 1 – Data Quality Categories and Dimensions

In our perspective, it is not possible to manage different data quality perspectives of each data user using an entity-relationship model.

To model the several dimensions of data quality requirements we propose instead the use of role modeling, within an object-oriented paradigm.

We shall use, in data quality modeling, two concepts, namely (1) *quality parameters* and (2) *quality indicators,* forming *the quality attribute,* (Wang, 1995).

Quality *parameters* relate to quality dimensions (qualitative aspects) while quality *indicators* relate with measurable attributes (quantitative aspects, normally from physical goods).

Quality parameters are related with subjective qualifiers (quality dimensions), while quality *indicators* are objective qualifiers (the quantitative aspects of data quality), i.e. the instantiation of quality dimensions.

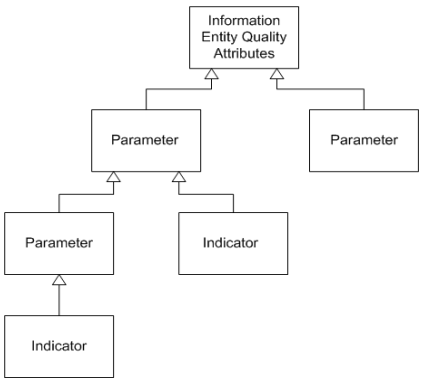

Figure 1 – Quality Attributes Meta-Model.

They diverge from each other only on semantics, but syntactically they are of the same nature. For example, as a data quality parameter we need to attain *confidence* and *timeliness*, and as a quality indicators we can decompose on *source data*, and *creation date.*

The quality parameters and indicators form the data quality *attribute.* This results in a hierarch tree that helps defining the data quality requirements, Figure 1.

Data quality requirements are related with information entities, representing business objects, The Business Object Management Special Interest Group (BOMSIG) defines it as "A representation of a thing that is active in the business domain, including at least its business name and definition, attributes, behavior, relationships, and constraints. A business object may represent, for example, a person, place, or concept. The representation may be in a natural language, a modeling language, or a programming language", (BOMSIG, 1995).

Attributes are facts about the business object relevant to fulfilling its business purpose, and are defined as the basic unit of information about any entity occurrence. They may be local or intrinsic to the entity, like (Name, Description) or inherited by relationship from another entity.

Quality attributes are multi-dimensional and hierarchic concepts, which provide a comprehension base to understand the characteristics that define data quality.

3. PROPOSAL

The problem we intend to solve consists in how to guarantee that attribute values of a business entity are correct and correspond to the defined semantics. By correct we mean that data is accurate syntactically, according the data users quality definitions, and semantically, according the established meaning, as weight is really weight and not height, for instance.

The semantics depend on the context where the model is applied, to gather the organizational meaning on problem resolution. To define data quality requirements, we need multiple dimensions, depending on specific needs from different organizational levels. For instance, the sales division may need inventory data to be accurate and complete, while management may need information that gathers other data quality dimensions, like reputation or timeliness of data for the decision making process. We might have different needs for the same data on the database or data warehouse systems, and there might be overlapping data quality dimensions for the same data, but modeling must be seen on different levels, according the data users.

The first problem we aim solving is how to relate quality management with data quality.

The second problem we aim solving is how to relate data quality at different organizational dimensions.

3.1 Business Process Modeling Approach towards Data Quality

In this paper we use an object-oriented business process modeling framework. This framework to represent the interaction between process activities, business goals, resources and information systems (Vasconcelos, 2001).

To solve the first problem stated above, we propose a business process pattern to ensure data quality in an organization. To solve the second problem, we define data quality attributes upon information entities having different meanings depending on the business view, Figure 2.

Business processes are of two natures, regarding organization's value chain, (Porter, 1985): Core Business Processes of primary activities that produce value within the organization; Support Business Processes that represent the support activities of an organization.

In this paper we introduce the notion of *Tertiary Organizational Processes*. Tertiary processes represent activities that cross over operational and support processes planes, interacting with the active entities within, with the purpose of achieving some special purpose objective. The modeling of tertiary processes is often facilitated by the introduction of the concept of an "orthogonal" plane, relative to the operational and support business process planes, where it is formed, on a need basis, to achieve special purpose objectives, generally managed as a project within the organization. We shall now show how to use the Tertiary Business Process Plane to assure data quality in the operational (core and support) plane.

We so forth consider three layers in which lies the organization modeling of any operational business process. Figure 2, depicts the main conceptual structure and layer separation we use on our approach.

The first layer models the operational activities, whereas the third layer models the data quality activities. The middle layer deals with data modeling, where designed classes of objects "resources" are instantiated from both perspectives, the operational *versus* quality (tertiary) processes.

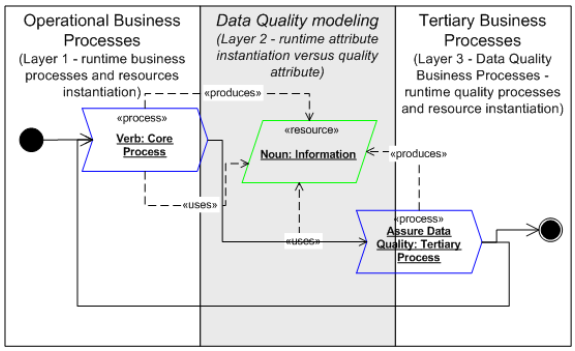

Figure 2 – Process diagram on business processes and data quality interaction.

3.1.1 The Resource Stereotype

In business process modeling, resources are objects within the business that are manipulated by processes. Resources may be arranged in structures, have relations with each other, and may be produced, consumed, used or refined in processes, (Eriksson, 2000). The class notation for resources used with the UML stereotype is «resource». Resources can be specialized into classes , such as information, thing, physical, abstract and people.

The business object in discussion is the resource stereotype, which we focus in the informational *entity* as specialization of a resource class.

A resource is an entity which can play a role in the realization of a certain class of tasks, (Vernadat, 1996) and a resource is also defined as a concept used in the business, and represents anything that we choose to evaluate as a whole, (Darnton, 1997).

The business and information systems modeling techniques differ on the business representation perspective (Curtis, 1992). There are four organizational perspectives: functional, behavioral, organizational and informational. The last one makes use of information entities representation, i.e., data produced or consumed by a business process.

The structure of an information entity structure or data model, Table 2, displays which attributes are normally used in the requirements survey for business process modeling.

The quality parameter is therefore an information entity attribute, representing the qualitative or subjective dimension of data quality, e.g. required data believability or timeliness.

Therefore, the data user defines the parameter and indicator values direct or indirectly. This is accomplished by indicating the quality characteristics needed on parameters and by measuring and filling the indicators values.

Entity Name	**{entity name}**	**Entity n.**	**{n}**
Sub-Entities	{sub-entities list}		
Alternative Name	{name}		
Identifier	{n}		
Type	{thing, people, abstract, information}		
Description	Textual description of entity and its utility and application		
Relations	Existing relations with other Entities (n) or Sub-Entities (n.n)		
Quality Parameters	Required data quality dimensions.		

Table 2 – An information entity structure template regarding the quality attributes.

Table 3 shows the information sub-entity structure, whose hierarchic relationship depends of indexation and presence of a corresponding textual description. In addition, in table 3, the information sub-entity structure includes the quality indicator attribute, referring to its quantitative data characteristics, such as *data source*, and *creation date*. Quantitative data must have a well-defined unit of measurement.

Sub-Entity Name	**{entity name}**	**Sub-Entity n.**	**{n.n}**
Alternative Name	{name}		
Type	{thing, people, abstract, information}		
Description	Textual description of entity and its utility and application		
Relations	Existing relations with other Entities (n) or Sub-Entities (n.n)		
Quality Indicators	Required data quality indicators related with entities quality dimensions.		

Table 3 – Information sub-entity structure template.

There is inherent complexity in the information sub-entity structure because, for the same data, we have overlapping and crosscutting requirements between core business processes, support business processes and tertiary processes –such as the data quality processes, shown in Figure 3. The attributes and methods of an information entity are enacted according to the assumed perspective of the respective plane of existence of the enacting invocation, at low granularity level.

We shall also consider predefined quality classes, with the data quality attributes of conceived business objects. These classes compose the quality attributes into the information entity. *We are thus extending the concept of information entity by using*

new attributes devoted to quality, beyond the usual and basic pre-defined data attributes.

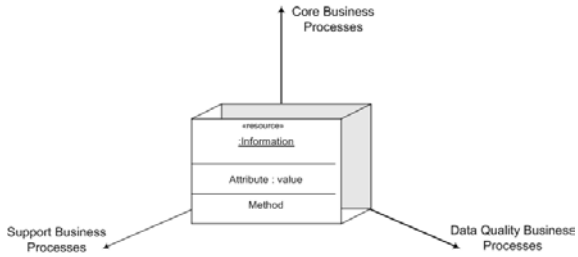

Figure 3 – Information Entity on 3 Dimension Process Perspectives

Since quality indicators help evaluate the trust character of the data, we propose that information entities should also contain an additional attribute for operational (current or historical data, called *primitive* data) and a second attribute, for qualitative data indication (or *derived* data), for further analysis, (Inmon, 1999). Data may also be characterized as *public* or *private*, but their application scope is, at this time, not our primary focus of interest.

Given that zero defects might not be reachable, or might be unnecessary or would have prohibitive costs, it is rather useful to evaluate data quality without the need for full inspection or regeneration of data.

Regarding data quality, we consider in our proposal two levels of granularity: 1) at high level when focused in a data quality business processes pattern, as discusses in 3.1.2; 2) at a low level, when focused on the pattern activities execution upon information entities, consumed or produced by business processes, as discusses in 3.1.3.

Our research does not focus at this time on software tools, data quality algorithms or even statistical control, although they might be considered implicitly when necessary.

3.1.2 Data Quality Process Pattern

We propose a data quality pattern at business process level. The pattern makes integrates two best practices from the quality management universe, namely English's (1999) information quality and Huang's (1998) data quality management. The pattern consists on a business process model that can be reused through adaptation in specific organizational scenarios. Figure 4 depicts the flow between pattern's top activities. The details on the interaction with entities are left to the next levels of functional decomposition. It is important to note that pattern deals with tertiary process activities pattern and applies to a orthogonal organization plan, as

previously discussed on section 3.1.

Figure 4 – A data quality business processes pattern

The "Quality and Architecture Definition" and "Evaluate Data Quality" activities focus on the evaluation of data quality, describing how the organization deals with data quality and what processes and resources are involved. The interaction between entities or resources, either consumed or produced by these activities, is only represented at lower granularity level to make possible activity autonomy and promote pattern reuse.

3.2 Data Quality Modeling

Data quality modeling does not depend on business process modeling. Data quality attributes should be tagged to business objects during entity/resource identification and definition.

The data quality processes relate with operational business processes by the manipulation of shared information entities (v. Figure 2).

3.2.1 Information Entities

The resource types are represented as classes while resource instances are represented as objects. Stereotypes to indicate different categories of resource types (v. Figure 5).

An information entity is a person, place, physical thing or concept that has meaning on business context and upon which is possible to keep information. Cook classifies an information entity as (1996) people or organization, place, thing, concept or event, i.e., non-physical or irrelevant part of any thing like a contract or invoice.

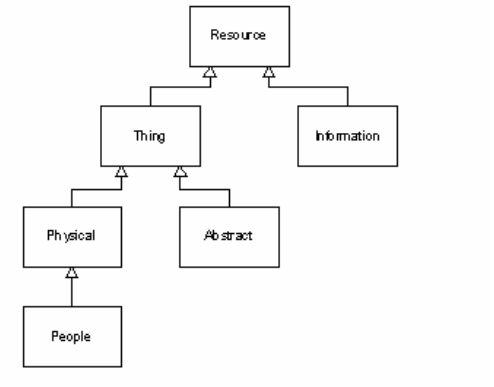

Figure 5 – Resources type hierarchic meta-model.

The information object is a representation of a concept, thing or another information object and it holds information that represents facts or knowledge about other resources (e.g. physical measuring of an object), or substitutes them. While the information object holds data, that does not make it the thing or concept it represents, (Eriksson, 2000).

3.2.2 Interaction with Information Entities

The information entity is modeled at two different levels: business process and class level.

The business process level is where the information entity is modeled as a resource stereotype and where it interacts with a process. The class level is where data is modeled in object-oriented classes, and where its attributes are specified.

Two processes may interact with shared information entities, like in a process interaction pattern. However, this pattern is only one-dimensional. However, to capture the information entity interaction in different contexts, we propose using role modeling concepts (v. section 3.2.3).

The data operations involving information entities and activities are CRUD operations, i.e., create, read, update and delete. Dashed lines represent these operations as in Figure 6, and correspond to object flow between resources and processes. These operations can mapped into a CRUD matrix. This matrix aligns activities and information entities. The dynamics in business processes is centered on the CRUD operations execution upon information (data) entities.

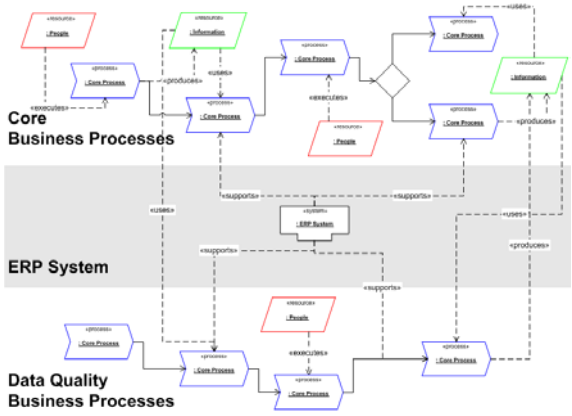

Figure 6 – Business processes interaction with information entities.

3.2.3 Role-based Modeling

A role represents some unit of responsibility or behavior. Actors play different roles while performing business process activities. Roles can be considered types in the sense they describe the behavior that is carried out by an instance of that role by a specific actor. Therefore, there may be multiple instances of the same role when a process is enacted; for instance, multiple customer roles may coexist when a sell process is being executed. A single actor may also play multiple roles. Although roles are independent of other roles, they communicate and are coordinated through interactions. Role models can be instantiated, aggregated and generalized.

In process modeling grouping activities according to roles improves the understanding of how responsibilities are set and how process activities are operationally carried out in short, it provides an insight of who is providing what to whom in the organization. This requires understanding the behavior of activities, resources and actors in an organizational context. This approach differs from current modeling approaches that focus on providing static representations of processes. By understanding the behavior of processes, we are providing the means to reuse it and adapt its organizational concepts.

3.2.4 Role modeling on data quality

Since role modeling allows the behavior of resources to be clearly separated and identified, we may have different data quality patterns since activities relate to resources according to these patterns of contextual usage.

As previously mentioned, a same resource is often used in multiple different contexts. Thus, we

can have different contexts of data quality as attributes of class that represents a business object. Resources are specialized so that its attributes and methods allow handling its quality features. Processes concerning quality attributes instantiate predefined classes and set values to its quality attributes. The quality attributes, using predefined combinations of quality parameters and indicators, allow judging the data quality.

This approach leads to a better understanding and confidence in data since quality information is kept within a resource (information entity type), which facilitates the data quality evaluation process.

4. CASE STUDY

This case study results from a research project on a real organization. This scenario illustrates the business processes of inbound logistics in a large warehouse. The targeted company handles an average of 22000 products and performs a few dozens of daily inbound transactions.

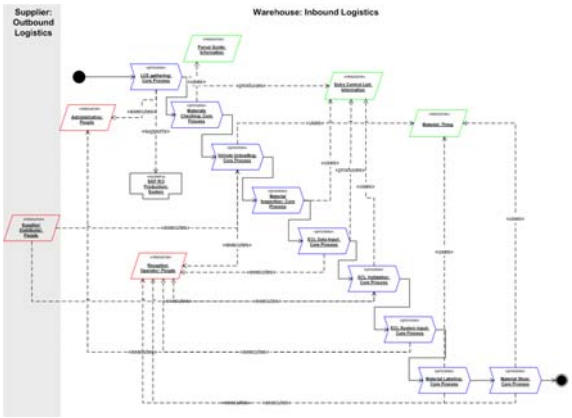

Figure 7 – Business Process Modeling of Inbound Logistics

Figure 7 depicts logistics business processes, which represent the materials incoming activities at the warehouse. The process starts when the materials arrive. The activities are material checking, material unloading, data input and ends with the material storing in the warehouse facility.

Figure 8 depicts how to ensure data quality at business process level making use of role modeling.

Regarding the "Data Quality Evaluation" (Figure 4) sub-process, it is represented in Figure 8, as "Data Quality Process", while the "Core Process" represents a logistics operational business process.

An oval represents the role associated to a class. It is a shorthand modeling for aggregating the role class with the base class.

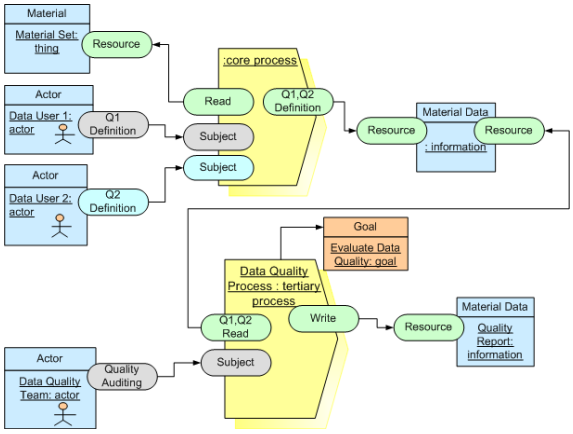

Figure 8 – Data quality evaluation using role modeling

In this example, "Material" corresponds to the "Material Set" used in the "Core Process". "Data Users" are the actors, which specify data quality requirements as part of a "Data Quality Definition" process. In this process, they define their respective data quality requirements, for instance: as Q1= {Timeliness, Completeness}; Q2= {Accuracy, Currency} for the "Material Set". The "Core Process" produces updated "Material Data" information entity, later audited using of a sampling set of materials. A "Data Quality Team" acts in the data quality requirements auditing. The "Data Quality Process" result is a "Quality Report" for later analysis. The overall process goal is to "Evaluate Data Quality" generated by "Core Process".

In this way, we have ensured data quality, using role modeling to manage different context and semantics of data quality, combined with the core or support business processes, which have data quality support in their information entities.

5. CONCLUSIONS

This paper proposes a data quality pattern to model the data quality intrinsic to business processes. This pattern can be used for data improvement on any organization, and makes use of a set of business processes to syntactically validate the data according to a model which depicts qualitative and quantitative data quality attributes. Role modeling is used to manage different quality contexts quality from different data users, assuring data quality at semantic level.

This contribution addresses data quality from a organizational engineering perspective using business process modeling, leveraging data

confidence and promoting continuous data quality improvement.

REFERENCES

Ballou, D.P., Pazer, H.L., 1987. Cost/Quality Tradeoffs for Control Procedures in Information Systems. International Journal of Management Science, 15(6), pp. 509-521.

BOMSIG, Business Object Management Special Interest Group, 1995. OMG Business Object Survey (Supply Side). OMG Document 95-6-4, OMG.

Curtis, W., Kellner, M., Over, J., 1992. Process Modeling. Communications of the ACM, 35, 9, pp.75-90, 1992.

Crosby, P.B., 1984. Data Quality Without Tears. McGraw-Hill.

English, L.P., 1999. Improving Data Warehouse and Business Information Quality, methods for reducing costs and increasing profits. Wiley.

Eriksson, H.E., Penker, M., 2000. Business Modeling with UML: Business Patterns at Work. OMG Press.

Huang, K., Lee, Y.W., Wang, R.Y., 1998. Quality Information and Knowledge Management. Prentice-Hall, pp.9-29.

Huh Y.U., 1990. Data Quality. Information and Software Technology, 32(8), pp. 559-565.

Inmon, W.H., 1999. Data Architecture – The Information Paradigm. QED Technical Publishing Group.

Laudon, K.C., 1986. Data Quality and Due Process in Large Interorganizational Record Systems. Communications of the ACM, 4-11.

Porter, M., 1985. Competitive Advantage. Free Press, New York.

Vasconcelos,A, Caetano,A., Neves,J., Sinogas, P., Mendes R., Tribolet, J., 2001. A Framework for Modeling Strategy, Business Processes and Information Systems. 5th IEEE International Conference on Enterprise Distributed Object Computing, EDOC 2001, IEEE Press. Seattle, USA, September.

Wang, R.Y., Reddy, M.P., Gupta, A., 1993a. An Object-Oriented Implementation of Quality Data Products, WITS.

Wang, R.Y., Kon, H.B., Madnick, S.E., 1993b. Data Quality Requirements Analysis and Modeling", Sloan School of Management, MIT.

Wang, R.Y., Reddy, M.P., Kon, H.B., 1995. Toward quality data: An attribute-based approach. Decision Support System 13, pp.349-372.

Wang, R.Y., Ziad, M., Lee, Y.W., 2001. Data Quality. Kluwer Academic Publishers, pp. 2-61.