Accepted for publication in the WITS-'93 Conference Proceedings, December 1993
Orlando, Florida

Modeling Data Quality and Context Through Extension of the ER Model

October 1993 TDQM-93-13

Steven Y. Tu
Richard Y. Wang

Total Data Quality Management (TDQM) Research Program
Room E53-320, Sloan School of Management
Massachusetts Institute of Technlogy
Cambridge, MA 02139 USA
Tel: 617-253-2656
Fax: 617-253-3321

Acknowledgments: Work reported herein has been supported, in part, by MIT's Total Data Quality Management (TDQM) Research Program, MIT's International Financial Services Research Center (IFSRC), Fujitsu Personal Systems, Inc. and Bull-HN. The authors wish to thank Peter Chen, Stuart Madnick, M.P. Reddy and Veda Storey for their comments on the earlier versions of this paper. Thanks are also due to Nancy Chin for her help in preparing this manuscript.

Modeling Data Quality and Context Through Extension of the ER Model

Abstract Capturing data quality and context semantics at the early stage of database design is a critical issue for both database researchers and practitioners. As in traditional database design, users' quality and context requirements should be represented at the conceptual level. This paper first generalizes the issues governing data quality and context semantics to be an *Interattribute-Relationship* problem, and then examines the feasibility of the Entity-Relationship (ER) model as a solution. Investigation on alternatives using the existing ER constructs reveals that it is necessary to extend the ER model. An extension called *Attribute-Relationship* (AR) is proposed to extend the view of having relationship at the attribute level. The attributes that involve in the relationship are termed *strong attribute* and *weak attribute*. In the AR extension, the dual roles of the strong and weak attributes and the embedded existence dependence between them are represented through an identifying relationship. Finally, a range of integrity constraints related to this extension are presented.

1. Introduction

Research on semantic data modeling suggests that more meanings about application data should be captured in the data model [Codd, 1979; Hull & King, 1987; Kim, 1989] . However, the capturing of the semantics is a never-ending task because it involves various dimensions and categories. Among them, the semantics concerning quality (e.g., accuracy and timeliness) and context (e.g., currency units and trade price status) of data have significant implications to users in the business community [Madnick, 1992] . Without these two types of semantics, the data of poor quality and mismatched context may lead to erroneous decisions.

Although previous research has, to a certain of extent, solved some of the issues [Siegel & Madnick, 1989; Wang & Madnick, 1990a; Wang & Madnick, 1990b; Siegel & Madnick, 1991a; Siegel & Madnick, 1991b; Sciore, Siegel, & Rosenthal, 1992; Wang & Reddy, 1992; Wang, Reddy, & Kon, 1992; Siegel, Sciore, & Rosenthal, 1993; Wang, Kon, & Madnick, 1993] , it is still not clear how quality and contextual data are interrelated with the application data at the conceptual level. Chen, who first proposed the ER model, also suggested in [Chen, 1993] that the conventional ER approach needs to be extended to incorporate the quality/context aspects, which is the objective of this research .

This paper is organized as follows: Background concepts are summarized in Section 2. In addition, the quality/context modeling issues are generalized and abstracted as the *Interattribute-Relationship problem* in this section. In Section 3, we offer some alternative solutions using existing concepts in the ER model, from which we argue for an extension to handle the Interattribute-Relationship problem. Section 4 presents our proposed solution - the *Attribute-Relationship* Extension. Section 5 first discusses structural and integrity constraint issues, and then the advantages of our approach in terms of these constraints. Finally in Section 6, we state some conclusions and future work.

2. Background Concepts

Chen defined an *entity* as a "thing" that can be distinctly identified, and a *relationship* as an association among entities [Chen, 1976] . For each entity, there are attributes that characterize the properties of concerns to users. Following this definition, an entity exists with existence independence unless it is a weak entity. This concept of "existence independence" is essential in modeling data quality and context.

Let A be an attribute associated with an entity E, then the quality or contextual semantics of A can be expressed by another set of attributes M = {M1, M2, ..., Mi}. For example, if A is the attribute Telephone# of an entity Employee, then M, the set of attributes indicating the quality of the value of Telephone#, might be {Creation_time, Collection_method}. As another example, if A is the attribute Stock_price of an entity Stock_report, then M, the set of attributes indicating the context of the value of Stock_price, might be {Currency, Trade_price_status}.

Although quality and context are concerned with two different aspects of data, they can be generalized to be the problem between the attributes A and M. At the conceptual level, we will call A *Strong Attribute* (SA) and M *Weak Attribute* (WA) hereafter. The reason of the naming will be presented in Section 4.

We make three observations related to SA and WA below. First, the derivation of SA and WA, given an entity E, is subject to user requirements. For example, different organizations, such as hospitals and restaurants, may have different concerns for the entity Customer. These concerns are reflected in their choices of SA and WA.

Second, the conceptual roles of SA and WA in users' views are distinct. The purpose of SA is to describe the characteristics of entities whereas the purpose of WA is to describe the characteristics of SA.

Furthermore, SA is associated with the ER model at the entity level whereas WA is associated with the ER model at the attribute level.

Third, there is an existence dependence relationship between SA and WA (WA depends on SA). For example, the context value {Currency = U.S. dollars, Trade_price_status = closing_price} exists and is meaningful only when it is associated with some stock price value such as "200.25". The second and third observations together lead to the following *Interattribute-Relationship Problem*:

How do we model SA and WA using the ER model?

This problem may appear to be trivial, but it is actually a difficult one. Although the relationship between SA and WA is analogous to the relationship between regular entity and weak entity in the ER model, the former can not be modeled in the same way as the latter. This is because both SA and WA draw values from their own *value sets* and in general do not hold the identifying properties of entities. The research question, then, is whether there is another feasible approach to model SA and WA using the existing ER constructs.

One approach that might be suggested to solve the Interattribute-Relationship problem is to simply connect SA and WA with lines to indicate their relationship, as shown in Figure 1. The problem with this approach is that a data model is not just a notational diagram. It should also provide clear semantics such that every construct in the data model can be unambiguously and formally expressed [Ng, 1981] . Therefore, using only lines to denote the connection between SA and WA is not sufficient; their corresponding semantics should also be explicitly defined.

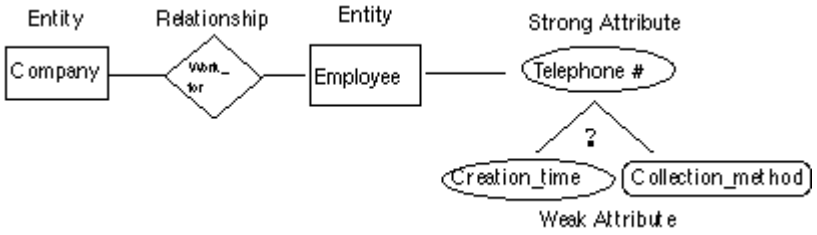

Figure 1: The Interattribute-Relationship Problem

3. Infeasibility in Using the Existing ER Constructs

In most traditional database design processes, after the users' requirements are elicited and before the logical data model (e.g., the relational model) is developed, the ER approach is adopted to derive a conceptual schema [Date, 1985; Elmasri, 1989; Teorey, 1990] . We consider this as a very important step in incorporating quality/context semantics into the database design processes. Since our scheme is to propose a solution for the Interattribute-Relationship problem after having identified the insufficiency of the current ER model, in this section we will examine various alternatives using the constructs provided by the original ER model. For each alternative below, we will point out its flaws and thereby disqualify it as a feasible solution. Then in Section 4, we will present an *Attribute-Relationship* extension to resolve the problem.

Using the existing ER constructs there are four mutually exclusive and collectively exhaustive alternatives that can be investigated for resolving the interattribute-relationship problem.

<Alternative 1>: As shown in Figure 2, the weak attributes can be attached to the entity Employee. However, three flaws exist in this alternative. First, since the weak attributes are treated as strong attributes

and associated at the entity level, the existence dependence relationship between SA and WA disappears. Second, since {Creation_time, Collection_method} is attached with Employee instead of Telephone#, it is not possible to distinguish whether {Creation_time, Collection_method} is to describe the attribute Telephone# or another attribute, say Employee_name. Third, conceptually the distinct roles of {Creation_time, Collection_method} and Telephone# are not explicitly shown in this alternative.

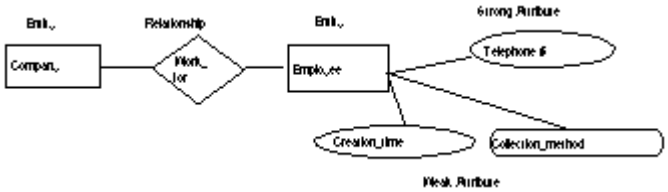

Figure 2: Attach the weak attribute to the entity

<Alternative 2>: As shown in Figure 3, the weak attributes can be attached with the entity Telephone#, which was originally the attribute of Employee. Again three flaws exist. First, based on Chen's definition, attribute Telephone# is a *value set* rather than an entity. Treating Telephone# as an entity will jeopardize the whole theory of the ER model. Second, if Telephone# is treated as an entity, then what is the relationship between Employee and Telephone#? It is inappropriate to treat this as a relationship between two entities (Employee and Telephone#) because the relationship between Telephone# and other entity attributes of Employee (e.g., Employee_name) will become undefined. Third, the relationship between the entity and SA is different from that between SA and WA. However, this difference is not shown in this alternative.

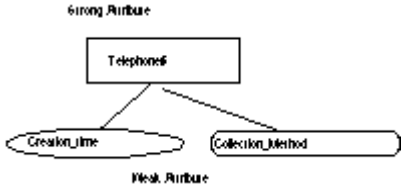

Figure 3: Change the strong attribute to entity

<Alternative 3>: As shown in Figure 4, SA and WA are connected by lines. As stated before, the flaw comes from the ambiguous and unclear semantics between SA and WA. This alternative may only serve as a simplified notation, but not as a solution for depicting the relationship between SA and WA in a data model.

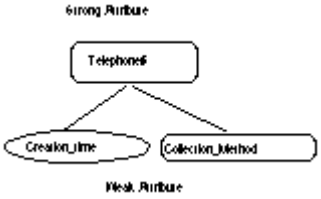

Figure 4: Connect SA and WA without providing semantics

<Alternative 4>: This alternative, as shown in Figure 5, seems to offer an appropriate solution. The connection between Telephone# and its WA is explicitly shown by the Quality_of relationship. Since both Telephone# and Tel_Q_Ent are considered as entities, this view satisfies the fundamental principle of the ER model that relationship is to associate entities. However, this alternative still suffers from the flaws of Alternative 2. Moreover, this alternative may only be feasible in the Object-oriented (OO) model scenario rather than in the ER model setting, and some of the underlying assumptions between the ER and OO models are not consistent. For example, in the OO model even a simple value can be considered as an "object", which however does not satisfy the definition of an entity in the ER model. A value, which has no

uniqueness independence in the real world, may be modeled as an "object" at the implementation level, but not as an "entity" at the conceptual level.

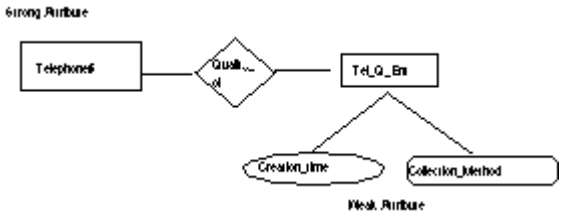

Figure 5: An object-oriented approach

4. An Attribute-Relationship Extension

From the above discussions, we conclude that the flaws exist because the ER model does not provide any mechanism to model the relationship at the attribute level. Therefore, in this section, we propose an Attribute-Relationship (AR) extension to incorporate the concept of relationship among attributes. In addition, we specify the constraints between the attributes explicitly just as the ER model does between entities. The AR extension we propose to resolve the Interattribute-Relationship problem is exemplified below.

As shown in Figure 6, Telephone# (SA) and {Creation_time, Collection_method} (WA) are still attributes. However, their relationship is described by the identifying relationship *Quality_of*. Furthermore, Telephone# is still an attribute of the entity Employee, except that it is also related to the weak attribute *Tel_Q_Att*, which is a composite *attribute* consisting of a set of weak attributes, {Creation_time, Collection_method} in this case. Telephone# is different from the other regular entity attributes, such as Employee_name. We elaborate on this extension below.

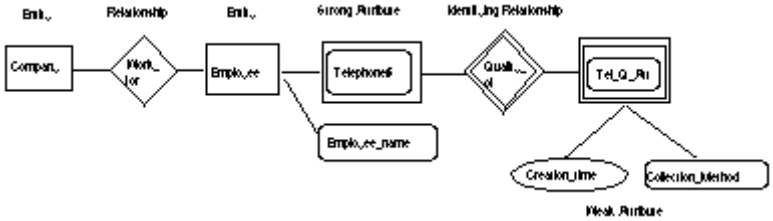

Figure 6: The attribute-relationship extension

1. Both WA and SA assume the dual roles of being an attribute and a virtual entity: In our AR extension, both Telephone# and Tel_Q_Att are still attributes. They do not have primary keys that identify their independent existence. At the same time, they are treated as virtual entities because they have a relationship analogous to that of strong and weak entities. Therefore, in Figure 6 SA Telephone# (SA) is depicted by an oval surrounded by a box while Tel_Q_Att (WA) is depicted by an oval surrounded by two boxes. The double boxes denote Tel_Q_Att as a *weak virtual entity*, much like weak entity in the ER model. This dual role offers the merits of incorporating the attribute-relationship extension and the backward compatibility with the entity-relationship concept.

2. The existence dependence is embedded in the identifying relationship: The existence dependence between SA and WA is expressed by the Quality_of relationship, as the double diamonds in Figure 6 depicts. This explicit representation of the Quality_of relationship enables us to further specify integrity constraints for SA and WA, much like those enforced in the existing ER model.

In passing, we make three observations: (1) Our extension to the ER model is minimal. The only concept invented is the relationship between attributes. Other concepts, such as weak entity and identifying relationship type, are adopted from the ER model. (2) Our extension is generalizable. For example, by replacing the Quality_of relationship with Context_of relationship, the WA will store the contextual semantics for Telephone#. (3) The diagrams proposed in our extension is self-explanatory.

5. Implications- Integrity Constraints Modeling

A good conceptual database design methodology, such as the ER model, should not only allow the structural aspects, but also the integrity constraint aspects as well to be expressed in the model. Under our model once the relationship between attributes is established, it is easy to enforce in the ER diagram the integrity constraints that should hold between SA and WA. This is very important in ensuring data consistency. In this section we will discuss a set of structural and integrity constraints. Through the discussion of these aspects, the advantages of our model over the previous work can then be appreciated. We explicate each constraint in the following:

1. Cardinality constraint: This constraint specifies the number of instances in SA that can be associated with their corresponding instances in WA. In general, the cardinality constraint is M:N.

N:1 Case As shown in Figure 7, suppose users require each stock price to be attached with only one context at a given time point. Suppose further that multiple stock prices can have the same context, then, we have a N:1 cardinality constraint between Stock_price and {Currency, Trade_price_status}.

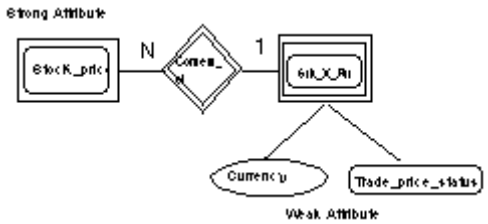

Figure 7: An N:1 cardinality case

1:1 Case As shown in Figure 8, suppose that two Telephone# values have distinct creation time and collection methods, then the cardinality constraint is 1:1

Note that this constraint comes directly from application requirements; it is not defined by the system.

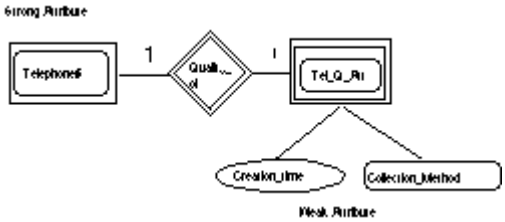

Figure 8: A 1:1 cardinality case

2. Participation constraint: This constraint specifies whether the existence of the participant of the relationship depends on the other corresponding participant. In our model, the participants are SA and WA that have an existence dependence relationship. Therefore, we can assert that the participation constraint

between SA and WA is Total (Figure 9), implying that every value in WA must be associated with at least one value in SA.

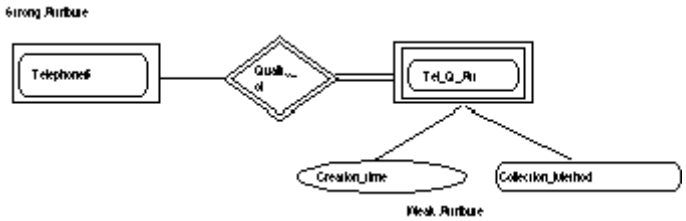

Figure 9: A total participant constraint

3. Non-key constraint: This constraint specifies whether there exists at least an identifying key. Although WA plays the dual roles of being an attribute and a virtual weak entity, there is no internal key associated with either role. The values in WA can be uniquely identified only by combining foreign keys from the "owner" entity of SA. Therefore, WA holds the property of non-key constraint.

4. Non-null constraint: This constraint specifies that whether null values are allowed or not. Users may require that for every value in the SA, its corresponding values in WA must exist. For example, users may require that for each value of Telephone#, the corresponding values for {Creation_time, Collection_method} must exist. In this case, the non-null constraint is enforced. Note that the non-null constraint for WA implies the total participation constraint for SA.

5. Cascading deletion/insertion/update constraint: This constraint specifies the legitimate states that should be satisfied after the deletion/insertion/update operations are performed to the database. Consider the scenario where an entity Stock_report has Company_name as its primary key, Stock_price as one of its SA's, and {Currency, Trade_price_status} as the corresponding WA. Table 1 exemplifies this constraint.

Table: Cascading constraints involving insert/delete/update operations

	Entity	SA (Stock_price)	WA {Currency, Trade_price_status}
Insert	insert a Stock_report	insert a value of "250"	insert values if there's a non-null constraint on WA
Delete	delete a Stock_report	delete the value "250"	delete the value { "U.S. dollar", "Closing_price"}
Update	update a Stock_report	update the value to be "120"	update the value to be {"Taiwan dollar", "Latest_price"}

6. Concluding Remarks

The capturing of data semantics is a never-ending task in data modeling. In this research, we focused on the semantics concerning quality and context of data. Without these two types of semantics, the data of poor quality and mismatched context may lead to erroneous decisions.

Although quality and context are concerned with two different aspects of data, they can be generalized to be the problem between the strong attribute (SA) and the weak attributes (WA) that describe SA. This is called the interattribute-relationship problem in this research.

By examining the possible alternatives, we showed that the interattribute-relationship problem can not be resolved using the existing ER constructs. We then proposed an attribute-relationship extension to the ER model. Central to the extension are the constructs of the identifying relationship Quality_of which links SA and WA (a composite attribute consisting of a set of weak attributes). The Quality_of relationship is depicted by double diamonds; the SA is depicted by an oval surrounded by a box; and WA is depicted by an oval surrounded by double boxes. Furthermore, the implications of the AR extension in terms of constraints modeling were explored.

Most previous work on ER extensions have focused on the concepts of generalization and specialization (supertype/subtype), object-orientation, function modeling, behavior modeling, user interface, and ER operators. This research differentiates itself from the others by extending the core ER model at the attribute level. We are currently conducting research to formalize the AR extension, and to investigate how the extended AR conceptual schema can be mapped to various logical models, such as the relational and object-oriented data models.

7. References

[1] Chen, P. P. (1976). The Entity-Relationship Model - Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1, 166-193.

[2] Chen, P. S. (1993). The Entity-Relationship Approach. In *Information Technology in Action: Trends and Perspectives.* (pp. 13-36). Englewood Cliffs: Prentice Hall.

[3] Codd, E. F. (1979). Extending the relational database model to capture more meaning. *ACM Transactions on Database Systems*, 4(4), 397-434.

[4] Date, C. J. (1985). *An Introduction to Database Systems.* Reading, MA: Addison-Wesley.

[5] Elmasri, R. (1989). *Fundamentals of Database Systems.* Reading, MA: The Benjamin/Cummings Publishing Co., Inc.

[6] Hull, R. & King, R. (1987). Semantic database modeling: survey, applications, and research issues. *ACM Computing Surveys*, 19(3), 201-260.

[7] Kim, W., Lochovsky, Frederick H. (1989). *Object-Oriented Concepts, Databases, and Applications.* New York: Addison-Wesley.

[8] Madnick, S. E. (1992). The Challenge to be Part of the Solution Instead of Being the Problem. V. C. Storey & A. B. Whinston (Ed.), In *Proceedings of the Second Annual Workshop Information Technology & Systems*, (pp. 1-9) Dallas, TX.

[9] Ng, P. (1981). Further Analysis of the Entity Relationship Approach to Database Design. *IEEE Transactions on Software Engineering*, 7(1), 85-99.

[10] Sciore, E., Siegel, M., & Rosenthal, A. (1992). Context Interchange Using Meta-Attributes. In *First International Conference on Information and Knowledge Management*, (pp. 377-386) Baltimore, MD.

[11] Siegel, M. & Madnick, S. (1989). Schema Integration Using Metadata. In Evanston, IL.: National Science Foundation.

[12] Siegel, M. & Madnick, S. (1991a). Context Interchange: Sharing the Meaning of Data. *SIGMOD Record, ACM Press*, 20(4), 77-79 (December).

[13] Siegel, M. & Madnick, S. E. (1991b). A metadata approach to resolving semantic conflicts. In *the proceedings of the 17th International Conference on Very Large Data Bases (VLDB),* (pp. 133-145) Barcelona, Spain.

[14] Siegel, M., Sciore, E., & Rosenthal, A. (1993). *Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems.* (No. 3543-93). MIT Sloan School of Management.

[15] Teorey, T. J. (1990). *Database Modeling and Design: The Entity-Relationship Approach.* San Mateo, CA : Morgan Kaufman Publisher.

[16] Wang, R. Y., Kon, H. B., & Madnick, S. E. (1993). Data Quality Requirements Analysis and Modeling. In *the Proceedings of the 9th International Conference on Data Engineering*, (pp. 670-677) Vienna: IEEE Computer Society Press.

[17] Wang, R. Y. & Reddy, M. P. (1992). *Quality Data Objects.* (No. TDQM-92-06). MIT Sloan School of Management.

[18] Wang, R. Y., Reddy, M. P., & Kon, H. B. (1992). Toward Quality Data: An Attribute-based Approach. *To appear in the Journal of Decision Support Systems (DSS).*

[19] Wang, Y. R. & Madnick, S. E. (1990a). A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. In *the Proceedings of the 16th International Conference on Very Large Data bases (VLDB)*, (pp. 519-538) Brisbane, Australia.

[20] Wang, Y. R. & Madnick, S. E. (1990b). A Source Tagging Theory for Heterogeneous Database Systems. In *International Conference on Information Systems*, (pp. 243-256) Copenhagen, Denmark.