Improving Data Quality Through Effective Use of Data Semantics

Stuart Madnick
Massachusetts Institute of Technology

***Abstract* – Data quality issues have taken on increasing importance in recent years. In our research, we have discovered that many "data quality" problems are actually "data misinterpretation" problems – that is, problems with data semantics. In this paper, we first illustrate some examples of these problems and then introduce a particular semantic problem that we call "corporate householding." We stress the importance of "context" to get the appropriate answer for each task. Then we propose an approach to handle these tasks using extensions to the COntext INterchange (COIN) technology for knowledge storage and knowledge processing.**

***Index Terms* – Data Quality, Data Semantics, Corporate Householding, COntext INterchange, Knowledge Management.**

I. Introduction

Data quality issues have taken on increasing importance in recent years. In our research, we have discovered that many "data quality" problems are actually "data misinterpretation" problems – that is, problems with data semantics. To illustrate how complex this can become, consider Fig 1. This data summarizes the P/E ratio for Daimler-Benz obtained from four different financial information sources – all obtained on the same day within minutes of each other. Note that the four sources gave radically different values for P/E ratio.

The obvious questions to ask are: "Which source is correct?" and "Why are the other sources wrong – i.e., of bad data quality?" The possibly surprising answer is: they are all correct!

Manuscript received November 3, 2003. Work reported herein has been supported, in part, by Banco Santander Central Hispano, Citibank, Defense Advanced Research Projects Agency (DARPA), D & B, Fleet Bank, FirstLogic, Merrill Lynch, MITRE Corp., MIT Total Data Quality Management (TDQM) Program, PricewaterhouseCoopers, Singapore-MIT Alliance (SMA), Suruga Bank, and USAF/Rome Laboratory.

Stuart Madnick is with the Massachusetts Institute of Technology, Sloan School of Management and School of Engineering, Cambridge, MA 02139 USA. (Phone: +1 617-253-6671; fax: +1 617-253-3321; e-mail: smadnick@mit.edu)

Source	P/E Ratio
ABC	11.6
Bloomberg	5.57
DBC	19.19
MarketGuide	7.46

Fig 1. Key Financials for Daimler- Benz

The issue is, what is really meant by "P/E ratio". Some of these sites even provide a glossary which gives a definition of such terms and they are very concise in saying something like "P/E ratio" is "the current stock price divided by the earnings". As it turns out, this does not really help us to explain the differences. The answer lies in the multiple interpretations and uses of the term "P/E ratio" in financial circles. It is for the entire year for some sources but for one source it is only for the last quarter. Even when it is for a full year, is it:

- the last four quarters?
- the last calendar year?
- the last fiscal year? or
- the last three historical quarters and the estimated current quarter (a popular usage)?

This can have serious consequences. Consider a financial trader that used DBC to get P/E ratio information yesterday and got 19.19. Today he used Bloomberg and got 5.57 (low P/E's usually indicate good bargains) – thinking that something wonderful had happened he might decide to buy many shares of Daimler-Benz today. In fact, nothing had actually changed, except for changing the source that he used. It would be natural for this trader (after possibly losing significant money due to this decision) to feel that he had encountered a data quality problem. We would argue that what appeared to be a data quality problem is actually a data misinterpretation problem.

To illustrate the significance of this issue, consider the vignettes displayed in Figs 2(a) and 2(b). In the case of Fig 2(a), the emissaries of the Austrian and Russian emperors thought that they had agreed on the battle being October 20th. What they had not agreed upon was which October 20th! This kind of semantic misunderstandings do not only resided hundred of years in the past, consider Fig

2(b) where a similar mishap also had dramatic consequences for the Mars Orbiter satellite.

(a) The 1805 Overture

In 1805, the Austrian and Russian Emperors agreed to join forces against Napoleon. The Russians promised that their forces would be in the field in Bavaria by **Oct. 20**. The Austrian staff planned its campaign based on that date in the **Gregorian calendar**. Russia, however, still used the ancient **Julian calendar**, which lagged 10 days behind. The calendar difference allowed Napoleon to surround Austrian General Mack's army at Ulm and force its surrender on Oct. 21, well before the Russian forces could reach him, ultimately setting the stage for Austerlitz.

Source: David Chandler, *The Campaigns of Napoleon*, New York: MacMillan 1966, pg. 390.

(b) The 1999 Overture

Unit-of-Measure mixup tied to loss of $125 Million Mars Orbiter

"NASA's Mars Climate Orbiter was lost because engineers did not make a simple conversion from English units to metric, an embarrassing lapse that sent the $125 million craft off course … The navigators [JPL] **assumed metric units** of force per second, or newtons. In fact, the numbers **were in pounds** of force per second as supplied by Lockheed Martin [the contractor]."

Source: Kathy Sawyer, *Boston Globe*, October 1, 1999, pg. 1.

Fig 2. Examples of consequences of misunderstood context

It should be apparent from these examples, and many more, that such "data quality" problems can have significant consequences. But in all these cases, the data source did not make any "error," the data that it provided was exactly the data that it intended to provide – it just did not have the meaning that the receiver expected.

Before going any further, it should be noted that if all sources and all receivers of data always had the exact same meanings, this problem would not occur. This is a desirable goal – one frequently sought through standardization efforts But these standardization are frequent unsuccessful for many reasons[1]. Consider Fig 3; is it a picture of an old lady or a young lady? The point here is that some will see it one way, some will see it the other way, and most be able to see both images – but only one at a time[2]. This is the situation that we often face in real life. There is often no "right" answer and different people will continue to see things in different ways. Merely saying that everyone should see it the same way does not change the reality that multiple different legitimate, and often essential, views exist.

[1] A full discussion of all the difficulties with standardization is beyond the scope of this paper. It is worth noting that the "Treaty of the Meter" committing the U.S. government to go metric was initially signed in 1875.

[2] If you are unable to see both, email me and I will send clues for seeing each.

Fig 3. Old woman or young woman?

II. Corporate Householding

In our research we have studied many examples of these "data quality" problems caused due to differences in data semantics. In this section we will introduce an interesting category of these problems, which we call the "corporate householding problem."

The rapidly changing business environment has witnessed widespread and rapid changes in corporate structure and corporate relationships. Regulations, deregulations, acquisitions, consolidations, mergers, spin-offs, strategic alliances, partnerships, joint ventures, new regional headquarters, new branches, bankruptcies, franchises … all these make understanding corporate relationships an intimidating job. Moreover, the same two corporation entities may relate to each other very differently when marketing is concerned than when auditing is concerned. That is, interpreting corporate structure and corporate relationships depends on the task at hand.

Lets us consider some typical, simple, but important questions that an organization, such as IBM or MIT, might have about their relationships:

[MIT]: "How much did we buy from IBM this year?"

[IBM]: "How much did we sell to MIT this year?"

The first question frequently arises in the Procurement and Purchasing departments of many companies, as well as at more strategic levels. The second question frequently arises in the Marketing departments of many companies and is often related to Customer Relationship Management (CRM) efforts, also at more strategic levels. Logically, one might expect that the answers to these two questions would be the same – but frequently they are not, furthermore one often gets multiple different answers for the same question even within each company.

These types of questions are not limited to manufacturers of physical goods, a financial services company, such as Merrill Lynch, might ask:

[Merrill Lynch]: "How much have we loaned to IBM?"

[IBM]: "How much do we owe Merrill Lynch?"

On the surface, these questions are likely to sound like both important and simple questions to be able to answer. In reality, there are many reasons why they are difficult and have multiple differing answers, as discussed in the next section.

A. A Typology of Corporate Householding Problems

At least three types of challenges must be overcome to answer questions such as the ones illustrated above: (1) identical entity instance identification, (2) entity aggregation, and (3) transparency of inter-entity relationships. These challenges provide a typology for understanding the Corporate Householding issues, as illustrated in Fig 4 and explained below.

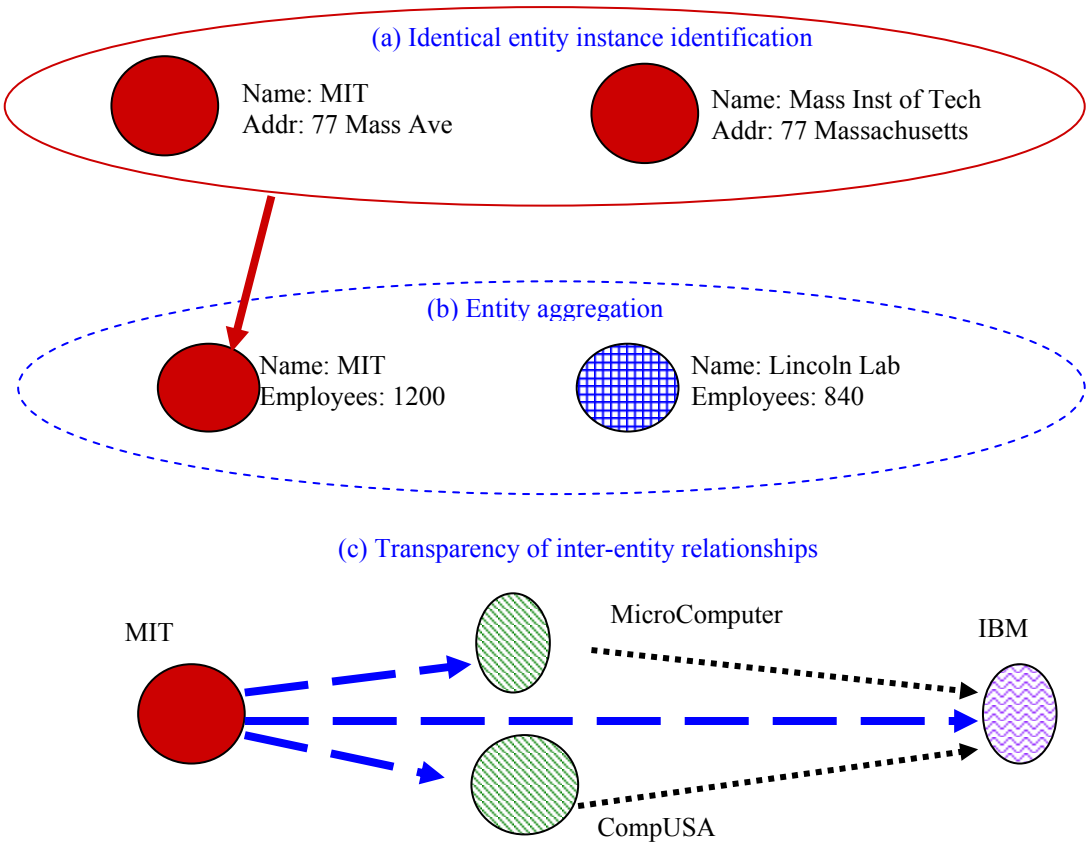
1) Identical entity instance identification. In general, there are rarely complete unambiguous universal identifiers for either people or companies. Two names may refer to the same physical entity even though they were not intended to create confusions in the beginning. For example, the names "James Jones", "J. Jones", and "Jim Jones" might appear in different databases, but actually be referring to the same person. Although identifiers such as Social Security numbers (SSN) are helpful, they might not always be available or feasible. For example, what is the SSN of a French citizen who works in one of IBM's European divisions?

The same problems exist for companies. As shown in Fig 4(a), the names "MIT", "Mass Inst of Tech", "Massachusetts Institute of Technology", and many other variations might all be used to refer to the exact same entity. They are different simply because the users of these names choose to do so. Thus, we need to be able to identify the same entity correctly and efficiently when naming confusion happens. We refer to this problem as *Identical Entity Instance Identification* [7]. That is, the same identical entity might appear as multiple instances (i.e., different forms) – but it is still the same entity.

2) Entity aggregation. Even after we have determined that "MIT", "Mass Inst of Tech", "Massachusetts Institute of Technology" all refer to the same entity, we need to determine what exactly is that entity? That is, what other unique entities are to be included or aggregated into the intended definition of "MIT." For example, the MIT Lincoln Lab, according to its home page, is "the Federally Funded Research and Development Center of the Massachusetts Institute of Technology." It is located in Lexington and physically separated from the main campus of MIT, sometimes refer to as the "on-campus MIT," which is in Cambridge. Lincoln Lab has a budget of about $500 million, which is about equal to the rest of MIT.

Problem arises when people ask questions such as "How many employees does MIT have?", "How much was MIT's budget last year?", or our original question – for IBM: "How much did we sell to MIT this year?" In the case illustrated in Fig 4(b), should the Lincoln Lab employees, budget, or sales be included in the "MIT" calculation and in which cases they should not be? Under some circumstances, the MIT Lincoln Lab number should be included whereas in other circumstances they should not

Fig 4. Typology of Corporate Householding



be. We refer to these differing circumstances as *contexts*. To know which case applies under each category of circumstances, we must know the context. We refer to this type of problem as *Entity Aggregation*.

3) Transparency of inter-entity relationships. A relationship between entities might involve multiple layers. Under what circumstances should these layers be collapsed? Let us consider our original questions again: [MIT] "How much did we buy from IBM this year?" and [IBM]: "How much did we sell to MIT this year?" As illustrated in Fig 4(c), MIT buys computers from IBM both directly and through local computer stores (e.g., MicroCenter and CompUSA). This is the classic case where a seller sells its products to a broker (and maybe directly also), and then the broker sells them to the ultimate buyer. Whether we are interested in the interface between the seller and the broker or the one between the seller and the ultimate buyer (via the broker) also depends upon the context – different answers will be appropriate for different contexts. We refer to this problem as *Transparency of Inter-Entity Relationships*.

B. Types of Entities and Their Relationships.

In considering the issue of entity aggregation, we need to consider what types of "corporate" entities exist and their relationships. There are obvious examples based on location (e.g., *branches*), scope (e.g., *divisions*), and ownership (e.g., *subsidiaries*). Even these may have variations, such as wholly-owned subsidiaries compared with fractional ownership – sometimes 66%, 51%, and 50% ownerships have different special significance regarding entity aggregation in matters of legal control, taxation, accounting, and bankruptcy.

In addition to these obvious types of entities, there are many others that need to be considered, such as *joint ventures*, which also might be fractional. Referring to our example in Fig 4, what type of entity is MIT's Lincoln Lab and how would one define its relationship with the other parts of MIT? Defining the "atoms" of corporate entities is an important part of our on-going research effort.

C. Wide Range of Corporate Householding Applications

There is a wide range of examples of Corporate Householding beyond the few examples used to illustrate the framework above. For example, if an agent is to determine a quote for business owner protection insurance for IBM, he must know how many employees IBM has [5]. To do so, he has to figure out what the rules are to decide what entities are part of IBM as far as business owner protection insurance is concerned. Does Lotus Development Corporation, a wholly-owned subsidiary of IBM, fall under the IBM umbrella? Similarly, if MIT buys a company-wide license for a piece of software, such as IBM's Lotus Notes or Microsoft's Office, does that automatically include Lincoln Lab – or not?

The concerns regarding Corporate Householding play an important role in both purchasing, and marketing activities. We have encountered many other specialized applications in discussing these matters with executives. For example, especially in consulting or auditing practices, you might agree with a client to not also do business with one of its competitors – but how is the "client" defined and how are its "competitors" defined?

III. Role of Context

We have used the term "context" earlier. To put this issue in perspective, consider a traditional family household. As family structures evolve, such as the increasing number of single families, families with no children, or husband and wife with different last name, it becomes more difficult to define and identify "household" [4]. For example, are grandparents or visiting cousins living at same address to be considered part of the same household? Are two unmarried people living together a household? The important point to note is that there is no single "right" answer; the answer depends upon the intended purpose of the question – which is what we mean by the context.

Similarly, a corporate household would also be different depending on different contexts such as a financial perspective, legal perspective, and the reporting structure. Identifying those contexts and representing the right structure for the right task is critical and can provide important competitive advantage.

Furthermore, it is important to note that corporate householding often changes over time; thus, the context also changes over time. For example, at one point Lotus Development Corporation was a separate corporation from IBM. When doing a historical comparison of growth or decline in "number of employees" of IBM, should current Lotus employees be counted in a total as of today? Should the Lotus employees in 1990, when it was a separate corporation, be added with the IBM employees of 1990 to make a meaningful comparison? Thus, temporal context often must be considered.

IV. Using Context Interchange (COIN) Technology for Storage and Processing of Corporate Householding Knowledge

COntext INterchange (COIN) [3] is a knowledge-based mediation technology that enables meaningful use of heterogeneous databases where there are semantic differences. For example, attributes that represent money, such as "price", may be expressed in "US dollars" in a USA database but in "Chinese RMB" in a Chinese database. Though the two attributes may have the same name, the semantic conflict has to be addressed before a correct query result involving the attributes can be obtained (e.g., "which price is less expensive?"). We refer to these semantic meanings as being the "context" of each source or

source context. Furthermore, different users, also called "receivers," may have different contexts or *receiver contexts* (e.g., I might want the answer in "Euros"). There are many parallels between the traditional COIN applications and the needs of Corporate Housekeeping where each source has its own Corporation Housekeeping context (e.g., "in this database, data on IBM includes all subsidiaries, such as Lotus") and each user's query has a context (e.g., "employee count for liability insurance purposes.")

The overall COIN project [3], [9], [10] includes not only the mediation infrastructure and services, but also wrapping technology and middleware services for accessing the source information and facilitating the integration of the mediated results into end-users applications. The wrappers are physical and logical gateways providing a uniform access to the disparate sources over the network [3].

The set of Context Mediation Services comprises a Context Mediator, a Query Optimizer, and a Query Executioner. The Context Mediator is in charge of the identification and resolution of potential semantic conflicts induced by a query. This automatic detection and reconciliation of conflicts present in different information sources is made possible by general knowledge of the underlying application domain, as well as informational content and implicit assumptions associated to the receivers and sources. These bodies of declarative knowledge are represented in the form of a domain model, a set of elevation axioms, and a set of context theories respectively. The result of the mediation is a mediated query. To retrieve the data from the disparate information sources, the mediated query is then transformed into a query execution plan, which is optimized, taking into account the topology of the network of sources and their capabilities. The plan is then executed to retrieve the data from the various sources; results are composed as a message, and sent to the receiver.

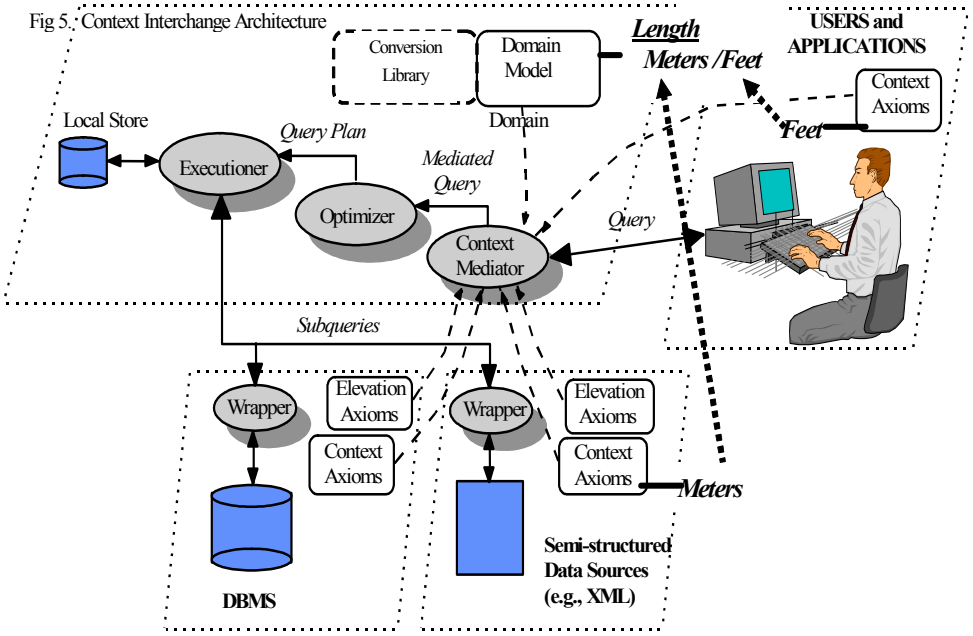
The COIN approach allows queries to the sources to be mediated, i.e., semantic conflicts to be identified and solved by a context mediator through comparison of contexts associated with the sources and receivers concerned by the queries. It only requires the minimum adoption of a common Domain Model, which defines the domain of discourse of the application.

The knowledge needed for integration is formally modeled in a COIN framework [3] as depicted in Fig 5. The COIN framework is a mathematical structure offering a sound foundation for the realization of the Context Interchange strategy. The COIN framework comprises a data model and a language, called COINL, of the Frame-Logic (F-Logic) family. The framework is used to define the different elements needed to implement the strategy in a given application:

- The Domain Model is a collection of rich types (semantic types) defining the domain of discourse for the integration strategy (e.g., "Length");
- Elevation Axioms for each source identify the semantic objects (instances of semantic types) corresponding to source data elements and define integrity constraints specifying general properties of the sources;
- Context Definitions define the different interpretations of the semantic objects in the different sources or from a receiver's point of view (e.g., "Length" might be expressed in "Feet" or "Meters").

Finally, there is a conversion library which provides conversion functions for each modifier to define the resolution of potential conflicts. The conversion functions can be defined in COINL or can use external services or external procedures. The relevant conversion functions are gathered and composed during mediation to resolve the conflicts. No global or exhaustive pair-wise definition of the conflict resolution procedures is needed.

Fig 5. Context Interchange Architecture



Both the query to be mediated and the COINL program are combined into a definite logic program (a set of Horn clauses) where the translation of the query is a goal. The mediation is performed by an abductive procedure which infers from the query and the COINL programs a reformulation of the initial query in the terms of the component sources. The abductive procedure makes use of the integrity constraints in a constraint propagation phase which has the effect of a semantic query optimization. For instance, logically inconsistent rewritten queries are rejected, rewritten queries containing redundant information are simplified, and rewritten queries are augmented with auxiliary information. The procedure itself is inspired by the Abductive Logic Programming framework and can be qualified as an abduction procedure. One of the main advantages of the abductive logic programming framework is the simplicity in which it can be used to formally combine and to implement features of query processing, semantic query optimization and constraint programming.

COIN was designed originally to address database-type[3] queries in the face of disparate semantics in different sources. We have recently adapted the COIN system so that it can be applied to corporate householding, which – in a certain sense – is to determine which attributes in different databases should be united or viewed as the same. In this implementation, the Domain Model stores general corporate householding knowledge. It decides how the relationships between entity instances should be decided when a certain task is concerned. The Elevation Axioms and Context Axioms describe the context associated with each specific database and specific application. The Context Mediator manages the interactions between Domain Model, Elevation Axioms, and Context Axioms. It is the interactions between the three that determine how the data stored in a database can be interpreted in terms of corporate household.

Such an implementation makes it much easier to answer questions such as "What is IBM's total global asset worth for purposes of bankruptcy insurance?", which involves both corporate householding knowledge processing and data semantics knowledge processing.

IV. Conclusions and Future Research

We are in the midst of exciting times – the opportunities to access and integrate diverse information sources, most especially the enormous number of sources provided over the web, are incredible but the challenges are considerable. It is sometimes said that we now have "more and more information that we know less and less about." This can lead to serious "data quality" problems caused due to improperly understood or used data semantics. The effective use of semantic metadata and context knowledge processing is needed to enable us to overcome the challenges described in this paper and more fully realize the opportunities. A particularly interesting aspect of the context mediation approach described is the use of context metadata to describe the expectations of the receiver as well as the semantics assumed by the sources.

In this paper, we presented a framework for understanding corporate householding problems. We then proposed that much of the burden of corporate householding could be reduced through the use of a corporate householding engine. We proposed an integrated method to accomplish the goal using COntext INterchange (COIN) to store and apply the captured knowledge. COIN builds on previous research, and is intended to maximally automate corporate householding with specially designed software modular – once the underlying source and receiver corporate household knowledge has been acquired.

Our future research plans include the following. First, we will continue to collect field data to determine the types of corporate householding knowledge needed. Second, we will explore the role of COIN in corporate householding. We plan to continue to extend our COIN-based system to further facilitate the process of capturing, storing, maintaining, and applying the corporate householding knowledge.

[3] COIN can process data in semi-structured web sites as if they were traditional relational databases using its "web wrapping" technology.

Acknowledgements

The participants in the MIT Summer Data Quality course (15.56s) and the MIT workshops on corporate householding are thanked for their helpful feedback and comments. Information about the Context Interchange project can be obtained at http://context2.mit.edu.

References

[1] Brown, J.S. and P. Duguid, *Organizational learning and communities of practice: toward a unified view of working, learning, and innovation.* Organization Science, 1991. 2(1): p. 40-57.

[2] Constant, D., L. Sproull, and S. Kiesler, *The kindness of strangers: The usefulness of electronic weak ties for technical advice.* Organizational Science, 1996. 7(2): p. 119-135.

[3] Goh, C.H., et al., *Context Interchange: New Features and Formalisms for the Intelligent Integration of Information.* ACM Transactions on Information Systems, 1999. 17(3): p. 270-293.

[4] Kotler, P., *Marketing Management: Analysis, Planning, Implementation, and Control*. 9th ed. 1997: Prentice Hall.

[5] Madnick, S., et al. *Corporate Household Data: Research Directions*. in AMCIS 2001. 2001. Boston, Massachusetts.

[6] Madnick, S., et al. *Improving the Quality of Corporate Household Data: Current Practices and Research Directions*. in Sixth International Conference on Information Quality. 2001. Cambridge, MA.

[7] Madnick, S. and R. Wang, *The Inter-Database Instance Identification Problem in Integrating Autonomous Systems*. in Fifth International Data Engineering Conference February 1989. Los Angeles, CA.

[8] Nonaka, I., *A Dynamic Theory of Organizational Knowledge Creation.* Organization Science, 1994. 5(1): p. 14-37.

[9] Siegel, M. and Madnick, S. *Context Interchange: Sharing the Meaning of Data*. SIGMOD RECORD, Vol. 20, No. 4, December 1991, p. 77-78.

[10] Siegel, M. and Madnick, S. (1991). *A metadata approach to solving semantic conflicts*. in Proc of the 17th International Conference on Very Large Data Bases, 1991, pp. 133-145.

Stuart E, Madnick (M'67) – is the John Norris Maguire Professor of Information Technology in the Sloan School of Management and a Professor of Engineering Systems in the School of Engineering at thr Massachusetts Instutute of Technology. Professor Stuart Madnick has been a faculty member at MIT since 1972. He has served as the head of MIT's Information Technologies Group for more than twenty years. He has also been an affiliate member of MIT's Laboratory for Computer Science, a member of the research advisory committee of the International Financial Services Research Center, and a member of the executive committee of the Center for Information Systems Research.

Dr. Madnick is the author or co-author of over 250 books, articles, or reports including the classic textbook, *Operating Systems*, and the book, *The Dynamics of Software Development*. He has also contributed chapters to books: *The Corporation of the 1990s: Information Technology and Organizational transformation* and *Information Technology in Action*.

His current research interests include connectivity among disparate distributed information systems, database technology, software project management, and the strategic use of information technology. He is presently co-Director of the PROductivity From Information Technology (PROFIT) Initiative and co-Heads the Total Data Quality Management (TDQM) research program. He has been active in industry, making significant contributions as a key designer and developer of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has served as a consultant to many major corporations, such as IBM, AT&T, and Citicorp. He has also been the founder or co-founder of several high-tech firms, including Intercomp, Mitrol, and Cambridge Institute for Information Systems, and currently operates a hotel in the 14th century Langley Castle in England.

Dr. Madnick has degrees in Electrical Engineering (B.S. and M.S.), Management (M.S.), and Computer Science (Ph.D.) from MIT. He has been a Visiting Professor at Harvard University, Nanyang Technological University (Singapore), University of Newcastle (England), and Technion (Israel), and Victoria University (Australia).