The Voice of the Customer: Innovative and Useful Research Directions

Dr. Stuart E. Madnick

John Norris Maguire Professor of Information Technology and
Leaders for Manufacturing Professor of Management Science
Sloan School of Management, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

The Need for a Reexamination

The book, *Computing the Future: A Broader Agenda for Computer Science and Engineering* by the National Research Council [HL92][1] starts its Executive Summary as follows:

> ... both the intellectual focus of academic CS&E [computer science and engineering] and the environment in which the field is embedded are today in the midst of significant change. ...

Further into the Executive Summary, the committee outlines its recommendations and priorities including:

> ... given the pressing economic and social needs of the nation and the changing environment for industry and academia, the committee believes that academic CS&E *must* broaden its self-concept or risk becoming increasingly irrelevant to computing practice.

In addition, the Association of Computing Machinery report on "The Scope and Directions of Computer Science: Computing, Applications, and Computational Science" [A91] advised:

> A close interaction between computer researchers and others is essential so that the questions under investigation remain connected to real concerns. Otherwise computing research can drift into irrelevance and cease to earn public support.

Speaking even more urgently, Professor Peter Denning, a former President of ACM, starts his article "Educating A New Engineer" [D92] with the observation that:

> University education is experiencing an enormous breakdown. An increasing number of students, employers, faculty, business executives, management specialists, public officials, and taxpayers have declared their dissatisfaction with the education and research that is available in most of our universities.

He proceeds to note "The connection between much of the research in universities and the concerns of people in business firms or general public concerns is not apparant either to the outside observer or to the researcher."

As one colleague noted, developing information-intensive solutions to complex global transportation management problems, such as intransit visibility, can be just as challenging research as developing new query optimization algorithms - but can have impacts that can be measured in billions of dollars of savings. Database researchers can either work proactively to define the future or they will be inevitably dragged into it.

The Challenge: Key Business Problems

Much of the content of this section emerged from this author's participation in the "Management in the 1990s" research project at MIT which culminated with the publication of the book, *The Corporation of the 1990s: Information Technology and Organizational Transformation* [M91, S91].

Business today is characterized by dramatically increased geographic scope due to globalization and world-wide competition, severe time and productivity pressures, and a rapidly

[1]This paper makes use of some material from that study and the author's participation in related panel discussions.

Proceedings of the 19th VLDB Conference
Dublin, Ireland, 1993

changing environment that often requires the restructuring of a company's operations in a very short time [M92, M93]. Movements toward consolidation (e.g., through bank mergers) are occuring at the same time as movements towards decentralization (e.g., the disintegration of IBM into multiple "Baby Blues").

Often the movement toward consolidation of previously independent companies or divisions into a single unit is motivated by the goal of gaining economy of scale. Thus, a rapid integration of separately maintained databases, processes, and networks would be necessary. Automated data conversion, interoperable network protocols, and transportable software systems are some of the major technological features necessary in such an environment.

Over and over the issue of time comes up. Popular phrases include "just-in-time", "continuous flow of information", "time-based functionality", and "time-to-market". Businesses are trying to compress the time from product concept to start of production, the time from product order to product ship, and the time to respond to a competitor's action. As an extreme case, many innovations in the financial services industry have a life span, from product concept to first deployment to final abandonment, of less than a month. In the business community, computer systems are often viewed as a major obstacle to time compression rather than a facilitator (one executive has a sign in his office that reads: "Once upon a time I thought that computers were the solution, now I realize that they are my major problem.")

One example that illustrates these needs and problems involves several ships full of goods being sent from the USA to help people in a foreign country suffering from a political disruption. While the ships were underway, the political situation improved in that country but a natural disaster struck another nearby country. The question posed was: Could one or more of the ships be diverted to help deal with the new disaster? This involved knowing what goods were supplied (they had come from multiple sources, each with their own computer systems), in which containers had those goods been stored, which containers were on which ships, and where was each of the ships currently located. Even though all of the necessary information existed in some computer system, there were so many disparate systems involved that a manual inspection of the contents of all the containers on each ship was ordered. Thus, a delay of several days occurred before the appropriate ships could be redeployed.

Although the above example may seem to be an interesting but isolated case, it represents much more the norm than the exception in dealing with complex global transportation situations. For example, during the Gulf War, of the 40,000 containers of material shipped to the Gulf, about 28,000 had to be manually unloaded and inspected in order to determine their contents. In general, the physical movement of material was faster than the movement of the supporting information. Transportation and logistic systems, in general, represent major challenges to the effective utilization of information technology [Q91].

Some Key Research Issues

Effectively integrating information from multiple sources both within organizations and between organizations represent both an important solution to many critical business needs [M91] and a key challenge for integration technology research [MSW90, MW91, SM89a]. An organization can be simultaneously "data rich" and "information poor" if they do not know how to identify, categorize, summarize, and organize the data. Although there are many important integration technology research needs in this area, four particular challenges will be highlighted as examples: client/server architectures, data semantics acquisition, data quality, and data semantics evolution.

Client/server control and evolution. The so-called "client/server" architecture is more than just technology, it represents an increased empowerment of the user and a dramatic increase in the decentralization of control. This has implications both for controlling the amount of load that is introduced by the users [GC93] and the ability to make use of and/or transition legacy systems [KF93].

Data semantics acquisition. As business operations become increasingly dispersed geographically and functionally, differences in work processes at each site performed by people trained for each site will become more critical, leading to data incompatibilities and inconsistencies when these differing sites must interact. For example, a certain insurance company has 17 different definitions of the term "net written premium" which is their primary measure of sales. Which definition is used depends on which office and functional group within the company is using the term and for what purposes the definition is being used. A useful integration of this company's databases would need to reconcile conflicting definitions of terms when necessary. Before these differences could be reconciled, we would need to be able to represent the semantics of the data as used in each environment, this is sometimes called the *context* of the data [SM91b]. Further research on using metadata to represent *context definitions* would provide the basis for capturing and disseminating knowledge about data meanings and facilitate the data reconciliation and integration process [SM89b, SM91a].

Data quality. Organizations have become very concerned about quality in areas ranging from manufacturing quality to software program quality. Data quality, in comparison, has received relatively little attention. Issues relating to data quality are becoming increasingly important as information is moved through the organization.[BP93] To a large extent, data quality considerations in the past were handled through personal familiarity; the user knew the characteristics of the data used in his or her organization and informally took this into account when using the data. This approach is not feasible as increasing numbers of information sources are used, many not well known to the user. They are increasingly exposed to data with various levels of quality for which they do not have first-hand familiarity. Furthermore, many currently automated processes for converting, merging, and manipulating the data renders inaccessible information about the original data that might have conveyed information about its quality. For example, the source of a given piece of information is often a key element in judgements about its credibility and quality [WM90]. There are many additional data quality attributes that may be important, such as accuracy, completeness, timeliness, and stability. Defining and measuring the important data quality attributes, which we refer to as *context characteristics*, and properly maintaining this quality-related information as data moves through and between systems represents a significant research challenge. With this quality information, decision makers would be better able to make more effective use of the data.

Evolving semantics. It must be realized that autonomous databases are independently evolving in semantics as well as in content (i.e., values). For example, consider the situation of stock exchanges around the world. Not only are the stock prices changing continuously, but the definition of the stock price also can change. At some time in the future, the Paris stock exchange will probably change from being measured in French francs to ECUs (European Currency Units). The normal "ticker tape" data feeds do not explicitly report the currency, it is implicit in the context of the source. More subtle examples include changes from reporting "latest nominal price" to "latest closing price" or from a percentage based pricing to actual prices, as currently happening at the Madrid stock exchange. Furthermore, in a historical database of stock prices, it must be recognized that the meanings had changed over time especially when doing a longitudinal analysis.

What is needed is not only a way to capture the meaning of each of these sources but also a way to represent the desired (or assumed) meaning of the receiver, which may be a human, an application, or another database. Then it would be possible to development a capability, which we refer to as a *context mediator*, to formally and automatically compare the semantics of the source and the receiver to determine if they are compatible, partially compatible, convertible, or incomparable. Research on the source/receiver model [SM91a] represents a direction towards solving the more general problem, which we call *context interchange* [SM91b].

Conclusions

It has been observed that the evolution of a field goes through periodic discontinuities. At these transition points, new paradigms and directions emerge. The VLDB community is uniquely positioned to be a leader in helping to make this happen. But, before we can help transform the field, we have to be prepared to transform ourselves.

Acknowledgements

Work reported herein has been supported, in part, by the National Science Foundation, MIT's International Financial Services Research Center (IFSRC), MIT's Leaders For Manufacturing (LFM) Program, MIT's PROductivity From Information Technology (PROFIT) Program, and MIT's Total Data Quality Management (TDQM) Program. A longer version of this paper is to appear in a forthcoming issue of *Decision Support Systems*.

REFERENCES

A91 Association for Computing Machinery, "The Scope and Directions of Computer Science: Computing, Applications, and Computational Science", *Communications of the ACM*, Volume 34, October 1991, p. 131.

BP93 Patterson, B., "The Need for Data Quality", *Proceedings of the VLDB Conference* (Dublin, Ireland), August 1993.

D92 Denning, Peter, "Educating A New Engineer", *Communications of the ACM*, Vol. 35, No. 12, December 1992, pp. 83-97.

GC93 Cohen, G., "A Practical Issue Concerrning Very Large Data Bases", *Proceedings of the VLDB Conference* (Dublin, Ireland), August 1993.

HL92 Hartmanis, J. and Lin, H, (Editors), *Computing the Future: A Broader Agenda for Computer Science and Engineering*, National Academy Press, Washington, D.C., 1992.

KF93 Fitzgerald, K., "Problems/Challenges facing Industry Data Base Users", *Proceedings of the VLDB Conference* (Dublin, Ireland), August 1993.

M91 Madnick, S., "Chapter 2: The Information Technology Platform", in *The Corporation of the 1990s: Information Technology and Organizational Transformation*, M. S. Scott-Morton (Editor), Oxford University Press, 1991.

M92 Madnick, S., "The Challenge: To Be Part of the Solution Instead of Being the Problem", *Proceedings of the Workshop on Information Technology and Systems* (WITS'92), December 1992, Dallas, Texas.

M93 Madnick, S., "Chapter 16: Putting IT All Together Before it Falls Apart", in *Information Technology in Action: Trends and Perspectives*, Richard Y. Wang (Editor), Prentice-Hall, 1993.

MSW 90 Madnick, S., Siegel, M. and Wang, Y.R., "The Composite Information Systems Laboratory (CISL) Project at MIT, *IEEE Data Engineering*, June 1990, pp. 10-15.

MW 91 Madnick, S. and Wang, R., "Logical Connectivity: Applications, Requirements, and an Architecture", *Proceedings of the 1991 Hawaii International Conference on Systems Sciences*.

Q91 Quiddington, P., "Cruising Along the Information Highway", *MIT Management Magazine*, Fall 1991.

S91 Scott-Morton, M. S., (Editor), *The Corporation of the 1990s: Information Technology and Organizational Transformation*, Oxford University Press, 1991.

SM 89a Siegel, M., Madnick, S. et al, "CISL: Composing Answers from Disparate Information Systems, *Proceedings of the 1989 NSF Workshop on Heterogeneous Databases*, December 1989, Evanston, IL.

SM 89b Siegel, M. and Madnick, S. , "Schema Integration Using Metadata", *Proceedings of the 1989 NSF Workshop on Heterogeneous Databases*, December 1989, Evanston, IL.

SM 91a Siegel, M. and Madnick, S., "A Metadata Approach to Resolving Semantic Conflicts", *Proceedings of the VLDB Conference* (Barcelona, Spain), September 1991.

SM 91b Siegel, M. and Madnick, S., "Context Interchange: Sharing the Meaning of Data", *SIGMOD Record*, December 1991, pp. 77-79.

US93 Schreier, U.,"Database Requirements for Knowledge-based Production Scheduling and Control: A CIM Perspective", *Proceedings of the VLDB Conference* (Dublin, Ireland), August 1993.

WM 90 Wang, R. and Madnick, S., "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective", *Proceedings of the VLDB Conference* (Brisbane, Australia), August 1990, pp. 519-538.