Data Quality Challenges in Enabling eBusiness Transformation

(Research in Progress)

Arie Segev, Professor & Director*
Fisher Center for Information Technology and Marketplace Transformation
University of California, Berkeley
http://haas.berkeley.edu/citm

Richard Wang

Associate Professor Boston University	Research Affiliate Fisher CITM, U. C. Berkeley	Co-Director for MIT TDQM Program

Abstract: This paper discusses data quality challenges in the context of eBusiness Transformation. It presents the major differences between traditional and eBusiness as they relate to business models, organizations, processes and technologies, and then outlines the differences with respect to data quality approaches. The scenarios described pose significant data quality (and other) challenges, and the paper discusses work in progress to construct a data quality strategy and implementation methodology.

1. INTRODUCTION

The field of data quality has witnessed significant advances over the last decade. Today, researchers and practitioners have moved beyond establishing data quality as a field to resolving data quality problems, which range from data quality definition, measurement, analysis, and improvement to tools, methods, and processes [1, 3, 5, 6, 11-19]. With many of the theoretical foundations developed, researchers have begun to go beyond the fundamental data quality research to solving critical business problems. For example, research has been initiated to investigate how to develop *data production maps* for information supply chain management and remanufacture [8]. Another area of active research is the conceptualization and software implementation for *corporate household* [10]. One research area that has not been actively pursued, however, is data quality in the context of eBusiness.

The Internet and eBusiness added new complexities to data quality primarily due the increase in a company's interaction with its environment – ***externalization***; and new levels of ***data integration*** resulting from new business models. That business-to-business (B2B) integration calls for the augmentation of data manufacturing models with ***data logistics*** concepts. Furthermore, it is imperative that organizations establish data quality strategies and implementation methodologies combined with their eBusiness transformation approaches.
In this paper we focus on B2B eBusiness, but there are obvious links to B2C eBusiness, for example, product and inventory information, which is used for B2C purposes, would inherit quality problems that were introduced in the data manufacturing process.

* The work of this author was supported by the External Acquisition Research Program (EARP) under contract N00244-99-C-0034

eBusiness Transformation entails business, organizational and technological aspects. It should be based on a comprehensive top-down view of the enterprise and its environment and incorporates proven principles when applicable. Basic principles of conventional information systems methodologies that have been developed in the last ten or more years still apply, but the scope and context have changed significantly. Section 2 discusses the eBusiness transformation process and elaborates on the business integration aspect. Section 3 then elaborates on the inter-company aspect and discusses four different scenarios; examples from the domain of B2B eProcurement are presented.

2. eBUSINESS TRANSFORMATION

eBusiness Transformation entails business, organizational and technological aspects. It should be based on a comprehensive top-down view of the enterprise and its environment and incorporates proven principles when applicable. Basic principles of conventional information systems methodologies that have been developed in the last ten or more years still apply, but the scope and context have changed significantly. The new context is characterized by:

- New business models, applications and related requirements
- The externalization level of companies
- The degree of required interconnectivity and integration
- The rate of change (technology and business models).

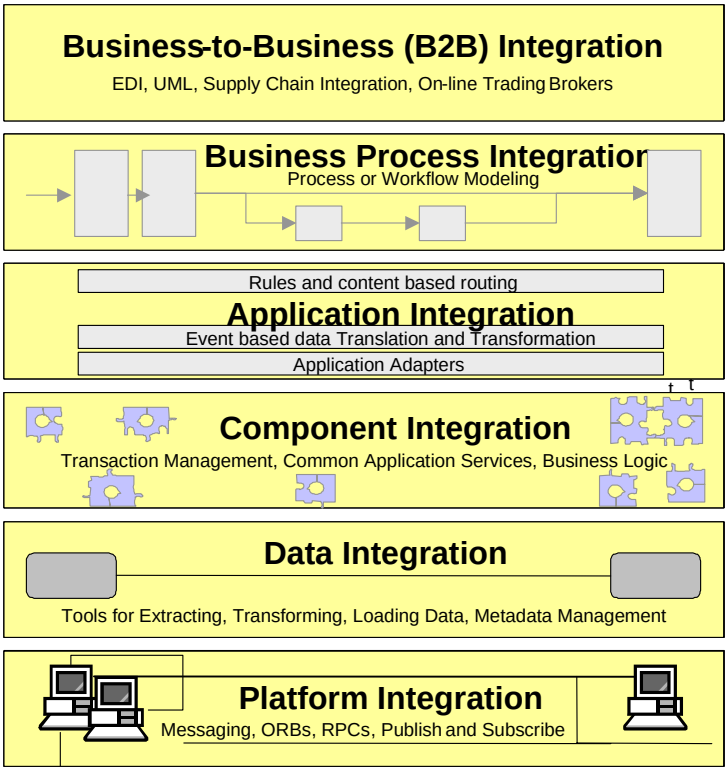

Figure 1: EAI Market Segmentation [Gold-Bernstein 1999]

The second bullet point indicates that increasingly, company's processes are shifted outwards as part of new business models involving interactions with customers, suppliers and partners. This, in turn, has led to an exponential increase of the company's interfaces (i.e., the level of business connectivity). From a process and data perspective a new level of Business-to-Business integration need emerged. A typical methodology used in addressing this need has been to expand the Enterprise Application Integration (EAI) technology beyond the corporate walls and delivers the full promise of eBusiness by integrating customers, suppliers and partners (see Figure 1). The basic principle is to create a decomposition-based application and technical infrastructure to support the business objectives and satisfy various performance constraints.

The previous figure represents sound decomposition principles, but it has the disadvantage of conveying a single company's perspective. We prefer to use the diagram of Figure 2, below to emphasize the new eBusiness requirements. It is important to note that while the requirements at the bottom of the figure are referred to as "technical," they of course have significant business and cost ramifications. The figure does not represent a decision model itself, but rather includes the scope, the elements, and the overall ideal order of the decision types. Frequently, one has to deal with a subset of the issues in a narrower and less systematic fashion, but whenever possible, this general framework should be followed or related to. The emphasize here is on the scope of the inter-organizational processes, the required infrastructure, as well as new organizational and skill dimensions.

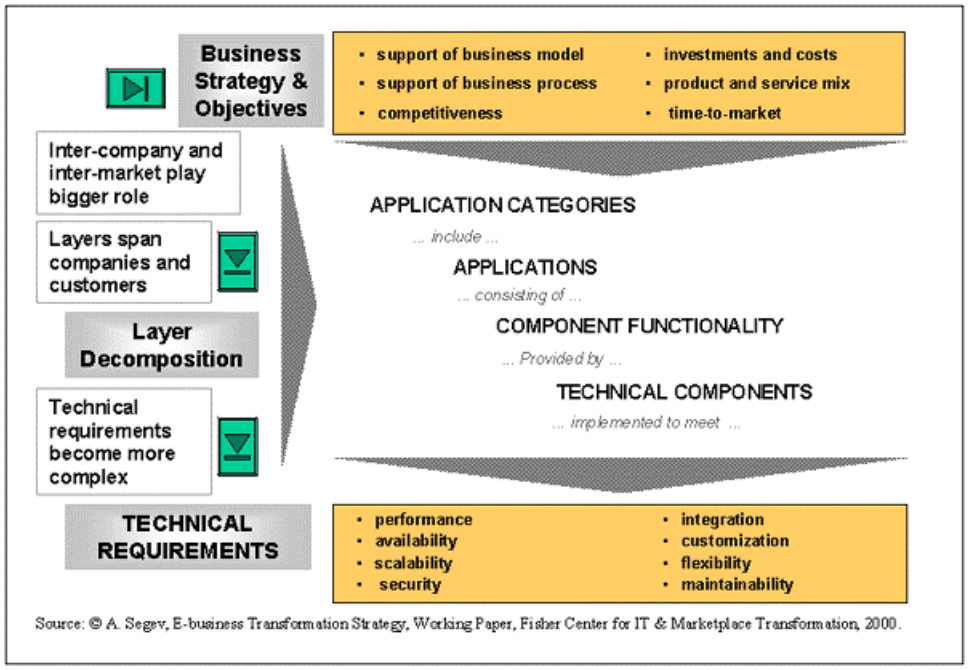

Figure 2: A Framework for eBusiness Integration

The new type of eBusiness applications involves a business and technology change in delivering products and services. An immediate requirement that companies face is to ***Web-enable*** legacy systems. Web servers, and the application server in particular, have become the foundation of business service delivery and, consequently, the Web service model must be central to

modernizing or moving out of legacy systems. It is important to understand that this model is as much a business model as it is a technology model. The range of Web-enabling possibilities is wide, but two general approaches are used when legacy systems are present. We refer to them as Level I and Level II. In Level I solution there is no significant change in functionality and it is based on creating an interface between the legacy system and the Web server. The concern here is the presentation and user interface, and it is similar to "GUI wrapping" that became a popular approach in the early days of client server. In the case of Level II solution, additional process functionality is introduced. The advantage of an application server is that it can be used for various degrees of Level I and Level II integration as shown in Figure 3. It also allows various degrees of inter-process integration and data quality enhancement. As an example, the application server enabled legacy application in the figure can provide data to a new object-oriented Web-based application, and the integration of the two at the application server provides the unified added value service that underlies the new eBusiness process. Furthermore, simple, but important, data quality enhancements can easily be introduced at Level II, e.g., performing validity checks on data attributes that were not implemented in the original legacy system. Cleaning the data at this junction is more effective and cheaper than data cleaning procedures downstream. In addition to the accuracy enhancement of Level I, Level II enhancements can be not only functional but also data quality. Relevant dimensions are completeness - enhanced through capturing more data and possibly relating it to other data (semantic completeness); timeliness – enhanced by capturing real-time data instead or in addition to other channels. In the next section we analyze in more details the data logistics as it moves across companies; the web-enablement approach described above is also applicable to many of those cases.

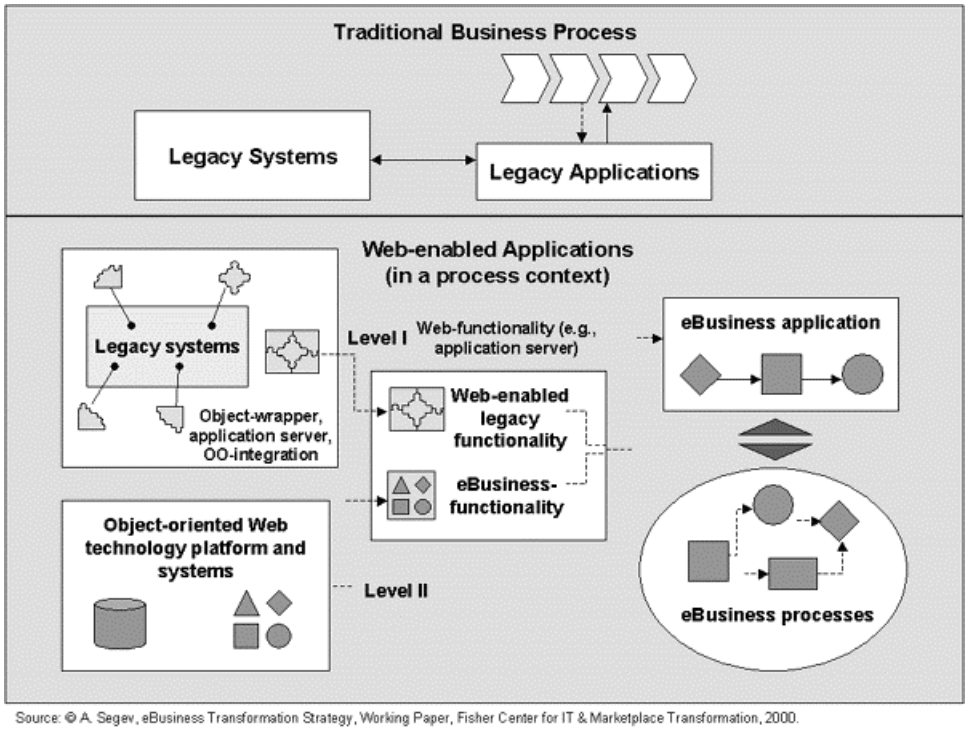

Figure 3: Web-Enabling Legacy Systems

3. INTER-COMPANY eBUSINESS INTEGRATION AND QUALITY ENHANCEMENT

There are four primary cases of inter-company eBusiness scenarios discussed below with respect to data quality strategies. These cases are discussed in the context of coordination and negotiation in [2], [9], [4], [7] .While not capturing all possible scenarios, we believe that these cases are the most important and represent the majority of real-life scenarios.

Case I: 1C

The case of a single company corresponds to the traditional intra-company data quality scenario. As discussed earlier in this paper, this case has received intensive attention in the last ten years both in academia and industry. The application server example in the preceding section is applicable to this case.

Case II: 2C

The case of two companies corresponds to dedicated systems between two trading partners, ranging from faxed papers and telephone calls, to traditional EDI and Web-EDI, to contemporary XML-based connectivity. Data quality issues in traditional systems (many of them are legacy systems) were identified and addressed long time ago both in research and in industry. In the case of fax, telephone errors are introduced due to "misunderstanding" and more errors through retyping,. There is often a "semantic reduction" as a result of translations to other systems. For electronic transmissions the following are common cases.

EDI: in addition to cost (setup and operational) many recipients print and re-input; in particular small companies. Translators and mappers improved situation somewhat. Further semantic problems arise in matching the received data with other data - from the same partner but from other systems, frequently arising because of the complexity, cost and time to modify existing EDI systems.

Web-EDI: primary objective was to reduce transport cost and possibly by-pass expensive VANs. One quality dimension improved is the timeliness when periodical downloads from VAN is replaced by more "real-time" web-based connectivity.

XML-based: These are contemporary systems, most implemented in the context of Case IV below. One should distinguish between two primary types of applications:
Transactional applications: including XML-EDI and new "pure" XML connectivity such as in Desktop Procurement Systems (DPS), e.g., DPS connectivity to inventory systems of the supplier. In many cases the 1-to-1 connectivity was changed to 3C2L by using the services as a content intermediary.
Collaborative applications: e.g. design, customer support; added connectivity and timeliness. More complete information. Workflow technology plays a major role.

Common problems are similar to those encountered twenty years ago when companies moved from file-based systems to databases by emulating the former on the latter, resulting in more efficient GIGO process. Unless this process is accompanied by a methodology-based process

and data quality improvement, the results will be similar, but with much more serious (and perhaps catastrophic) results to the business. The lower portion of Figure 4 illustrates the case of direct connectivity between the buyer and the seller in the context of eProcurement. It typically involves a significant business relationship that justified the cost of setting the one-to-one business integration. A main obstacle to data quality enhancement is the legacy EDI conduit which makes it difficult to add to the functional business integration, leading to parallel systems that don't integrate well relative to the end-to-end process. There is also typically ambiguity about the responsibility of each company for the data quality.

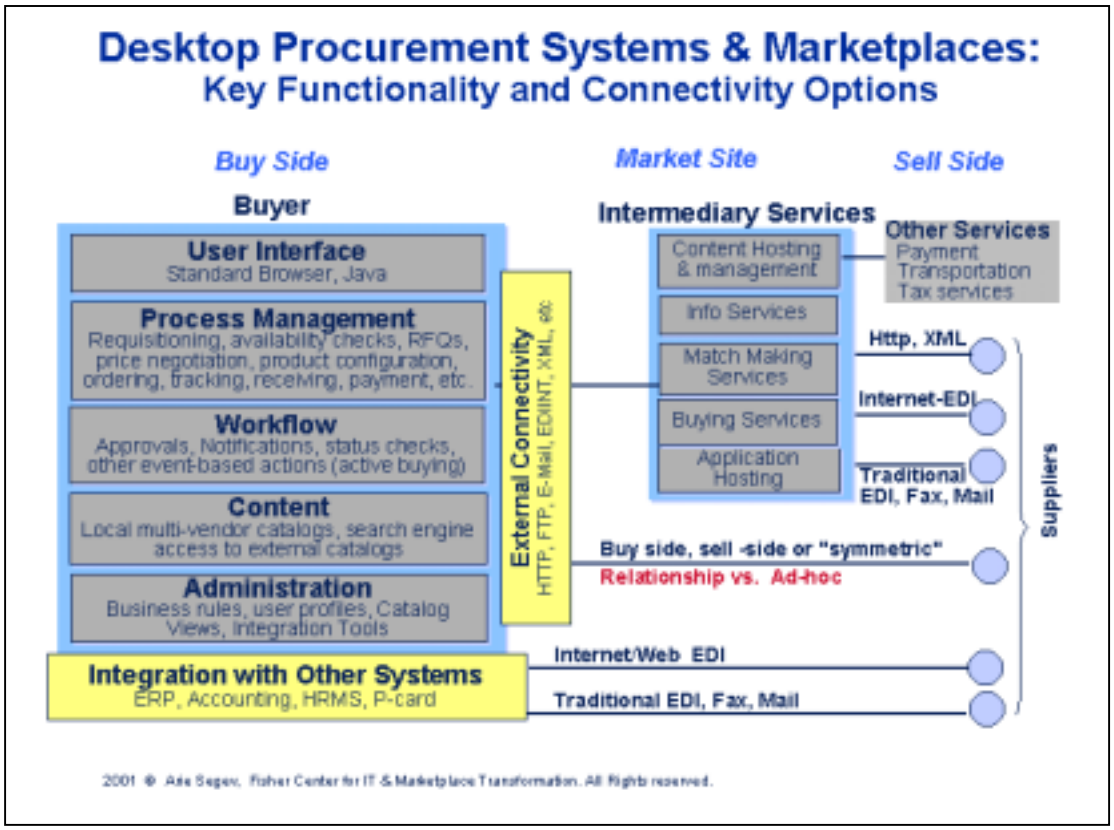

Figure 4: eProcurement Marketplace

Case III: nC1L

This case corresponds to providing a solution from multiple complementary vendors. Turn-key solution providers, e.g., restaurant kitchens, and travel packages are examples of traditional processes. Basic data quality problems in such an environment are not new and result form lack of collaboration among the suppliers, which are further complicated by severe data segregation constraints, e.g., a tour operator's data warehouse must maintain separate customer lists received from the various suppliers (airlines, hotels, etc). This last example was an information product example; for physical goods, there are various business models, including buying into inventory and then providing the components or serving as a broker and interacting with the multiple vendors.

In a dynamic eBusiness environment, major data quality problems have adverse effects on coordination, product compatibility, etc. The data quality problems are in integrating the data from the multiple vendors.

Case IV: nCmL

The case of n Companies and m Levels represents a supply web environment that includes elements of the previous cases. A particular sub-case, 3C2L, is the most common form of marketplace intermediation. The upper section of Figure 4, illustrates it in the context of eProcurement. Some of the most difficult data logistics problems arise here, since many suppliers are not enabled to move high quality XML-based data from their transactional systems to the e-catalogs of the intermediaries. The intermediaries add value by creating an integrated e-catalog after cleaning the data from the various sources, parts or whole of which are downloaded by the buyer. While the data semantics and completeness is generally better than the non-intermediated case, it can actually be less timely than the EDI solutions discussed in Case III.

The most general case involves multiple levels (supply chain) as well as multiple parties at a given level. It generally requires industry-wide standards for data and processes and often pursued in the interoperability context. Figure 5 represents such a case in the context of a joint project between the department of Architecture and CITM. It involves architectural and interior design processes and technologies combined with multi-level supply web environment, both B2C and B2B. Key functional components are listed in the figure, all present data quality issues. Document workflow systems are the basis of collaborative design and negotiations and they need to be integrated with e-catalogs and other pieces of information in many real-life situations. This introduces the most difficult interoperability issues, but without their solutions, the particular business model can't be implemented.

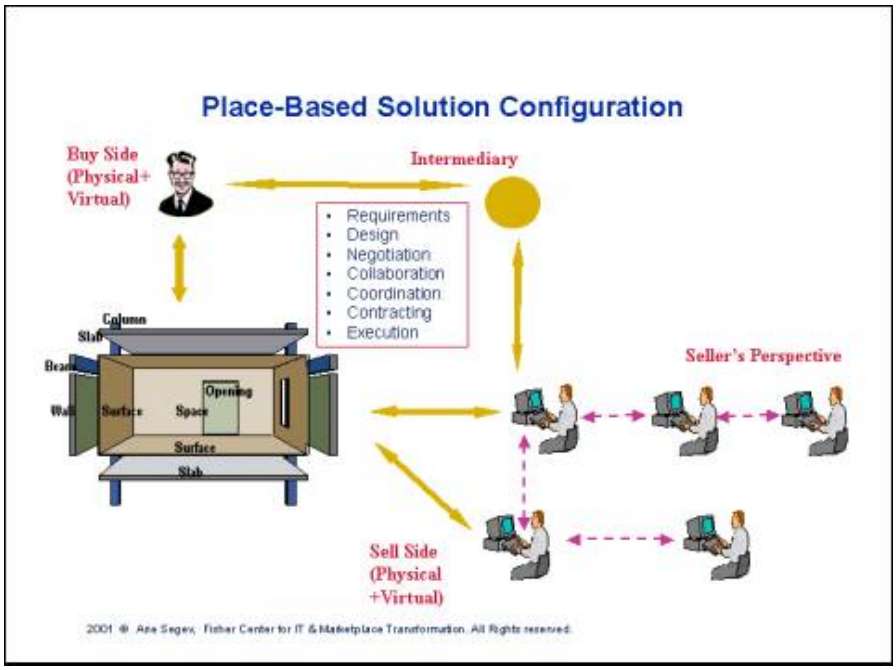

Figure 5: Place-based Solution Configuration

4. SUMMARY

This paper discussed data quality challenges in the context of eBusiness Transformation. The increased levels of company externalization make B2B integration a difficult proposition. That integration is of business models, processes, and technologies. The paper focused on eBusiness transformation for the B2B case and data quality implications. The eBusiness scenarios described pose significant data quality (and other) challenges; taxonomy of the scenarios and understanding the various data quality pitfalls are part of a data quality strategy designed to effectively deal with multiple points of data quality enhancement. The work presented in this paper is work in progress to construct a data quality strategy and implementation methodology.

REFERENCES

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

[2] Bichler, M. and S. A., Methodologies for the design of negotiation protocols for E-markets. *Computer Networks*, 2001.

[3] Fedorowicz, J. and Y. Lee, Accounting Information Quality. *Journal of Accounting Information Review*, *3*(1) 1999, pp. 1-7.

[4] Gold-Bernstein, B., EAI Market Segmentation. *EAI Journal*, 1999.

[5] Huang, K., Y. Lee and R. Wang, *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.

[6] Kahn, B. K., D. M. Strong and R. Y. Wang (1999). Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, Forthcoming, 2002.

[7] Kim, S. A. a. J., Frictionless Market and Automated Coordination. *CITM Working Paper*, 2001.

[8] Lee, Y., T. Allen and R. Wang. Information Products for Remanufacturing: Tracing the Repair of an Aircraft Fuel-Pump. in *Proceedings of Sixth International Conference on Information Quality*. Cambridge, MA: pp. 77-82, 2001.

[9] M., G., G. J. and S. A. Multi-Vendor Electronic Catalogs to Support Procurement: Current Practice and Future Directions. in *Proceedings of Bled Conference on Electronic Commerce*. Slovenia: pp. 1999.

[10] Madnick, S., R. Wang, F. Dravis and X. Chen. Improving the Quality of Corporate Household Data: Current Practices and Research Directions. in *Proceedings of Sixth International Conference on Information Quality*. Cambridge, MA: pp. 2001.

[11] Redman, T. C., ed. *Data Quality for the Information Age*. 1996, Artech House: Boston, MA. 303 pages.

[12] Storey, V. C. and R. Y. Wang. An Analysis of Quality Requirements in Database Design. in *Proceedings of the 1998 Conference on Information Quality*. Massachusetts Institute of Technology: pp. 64-87, 1998.

[13] Strong, D. M., Y. W. Lee and R. Y. Wang, Data Quality in Context. *Communications of the ACM*, *40*(5) 1997, pp. 103-110.

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

[15] Wang, R., J. Funk, Y. Lee and L. Pipino, *Journey to Data Quality*. MIT Press, Cambridge, Massachusetts, Forthcoming.

[16] Wang, R., M. Ziad and Y. Lee, *Data Quality*. Advances in Database Systems, ed. A. K. Elmagarmid. Kluwer Academic Publishers, Norwell, Massachusetts, 2001.

[17] Wang, R. Y., A Product Perspective on Total Data Quality Management. *Communications of the ACM*, *41*(2) 1998, pp. 58-65.

[18] Wang, R. Y., Y. L. Lee, L. Pipino and D. M. Strong, Manage Your Information as Product: The Keystone to Quality Information. *Sloan Management Review*, *forthcoming*, 1997.

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