On Information Quality and the WWW Impact
A Position Paper

Arie Segev
Fisher Center for Information Technology and Management
Haas School of Business
University of California, Berkeley
segev@haas.berkeley.edu

Abstract

This short position paper addresses some issues related to the practice of information quality improvement and the impact of the Internet and WWW related technologies on it. This impact is direct as well as indirect, positive as well as negative. The discussion also implies a general framework for devising quality improvements techniques and practices.

1. Information Quality Framework

The subject of information and data quality has been receiving significant attention in the last couple of years as a result of the increased importance and value of information for competitive. advantage . Major technological developments in the areas of communication, personal computing, databases, and CASE tools have led to both problems and opportunities. By making systems easier to develop, data easier to distribute and replicate, and by providing end-users with powerful desktop tools, data have proliferated in an "unmanaged" manner ; the same technologies, however, also provide the capability to improve data quality. In order to deal effectively with the issue of information quality it is important to:

1. Establish organizational awareness of the importance of information quality, and parties responsible for it.
2. define what we mean by data/information quality (i..e., quality dimensions or criteria)
3. establish information flows and processes map
4. identify quality problems (or potential ones) and their location on that map
5. identify technologies and practices that can be used to solve these problems
6. evaluate the cost/benefit tradeoffs associated with improving the quality of particular data or processes

The above points (they can also be viewed as steps to be taken) are quite intuitive but often not practiced. There has been significant work on defining quality dimensions (e.g., [6],[8]). While all the dimensions that have been

identified are important, some of them are not *intrinsic* properties of the data, e.g., believability or accessibility; this is not to say that those factors are unimportant, but that there are probably other mechanisms (rather than improving the quality of the data) that can improve them. It is important to separate the different aspects associated with data (such as intrinsic properties, and capture and delivery systems). The information flows and processes map is a good mechanism not only to do that, but also to separate causes from symptoms. A particular quality dimension may not be relevant in some contexts. For example, inconsistent data from multiple sources is generally considered a data quality problem where a common recommendation is to capture the data using a single source; this of course will not make sense in intelligence applications where data has to be gathered from multiple sources, and inconsistency does not necessarily mean *bad* data. Defining the quality dimensions explicitly is important both for communication and management purposes. Similar to information modeling we have to make sure that we mean the same thing when we use the term "data quality". The explicit dimensions facilitate the prioritization of actions taken to improve data quality as well as quality monitoring

Identifying the flow of data and the processes that effect it is extremely important as it helps in identifying those points where quality is susceptible to degradation, and where quality enhancement actions should take place. A very simplified diagram is introduced in Figure 1; D1 through D7 are common stages which data undergo; some may not be applicable for certain applications, and typically there is significant cycles which are not shown in the diagram (e.g., data in D6 is placed back in the database rather than displayed). .In practice a data item may go through a process type multiple times, e.g. it is captured, processed, distributed, processed again, distributed again, and so on. The diagram is not a system flow. The oval shapes indicate crucial steps that significantly effect data quality (with respect to all its definitions in the references); some of their instantiations are shown in Figure 2.

The quality dimensions are orthogonal to those processes. This enable one to examine a quality problem in a structured way by first defining the problematic quality aspects, separate symptoms from causes, and examine the cause where it occurs (any place in D1 through D7). For certain quality dimensions, the improvement points can be confined; for example, the quality dimensions of the conceptual view are impacted only in D1 (primarily) and D2 since they don't deal with the data instances. Data currency (or timeliness) for example can be impacted by several processes (D1 if it specifies currency requirements, and D3 through D6).

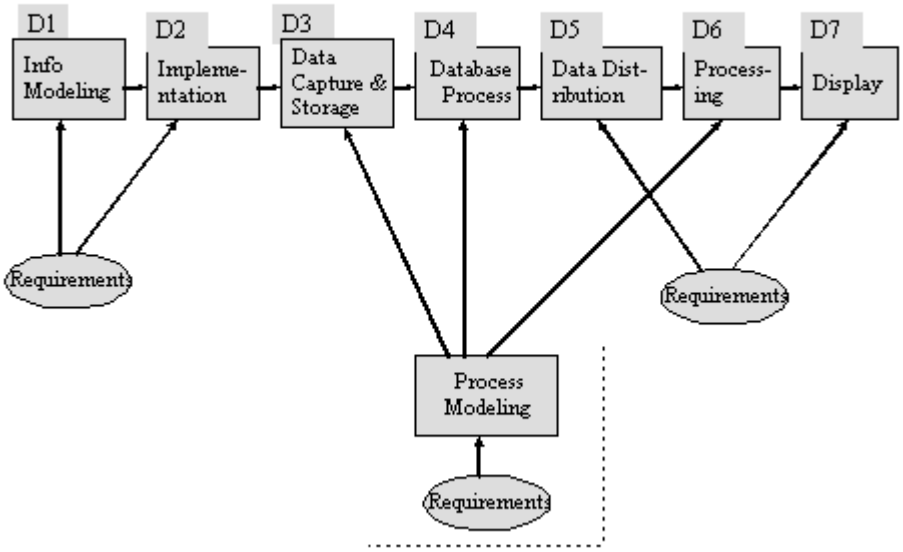

Figure 1: Quality Improvement Points

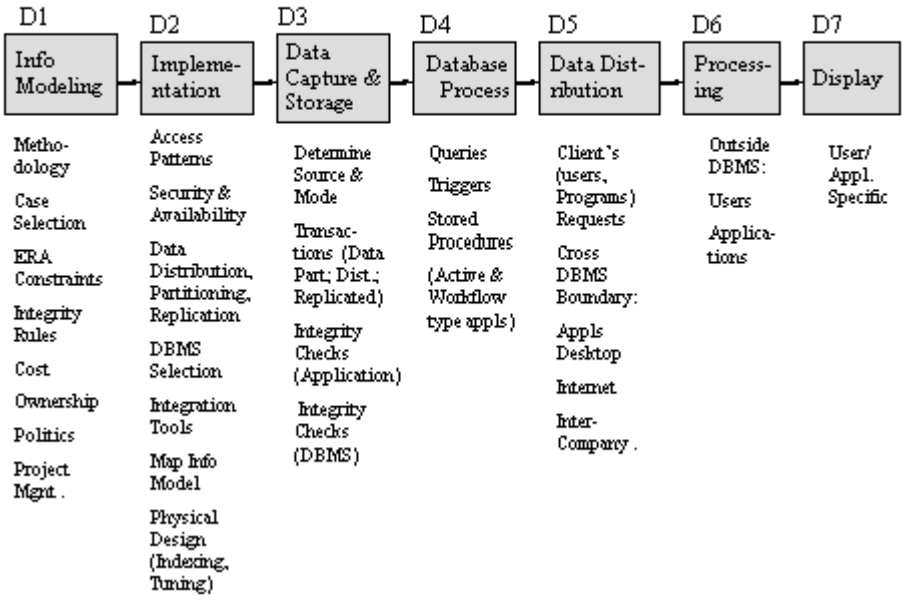

Figure 2: Quality Improvement Points (Detailed)

The above discussion and the diagrams assumed. a *managed data processes*. These processes determine the boundaries of the managed system. The big problem: data can leave and enter those boundaries in an unmanaged way. For example, data is extracted from a database into a spreadsheet, manipulated in various stages and then used to update the database; that loop may be subject to poor quality control. The problem is that the traditional lines of demarcation between users and IS have become increasingly blurred. The Internet and the WWW only makes this more prevalent and complicated (see discussion below

1. The Impact of the Internet/WWW on Information Flows

The advances in communication and computing technologies coupled with cross-functional business integration have changed information flows dramatically. Figure 3 illustrates the traditional information flows. Most of the organizational data was captures or generated by transaction systems, totally under the control of a centralized IS department, and then moved up the organization in the forms of reports (typically aggregation and classification of the lower level data). External data, about the business environment, entered the organization at a

higher (decision making) level. Conceptually, identifying the quality control points was relatively easy (though often not practiced): for external data, the only choice was source of data, and for internal data, the transaction systems, databases, and the report- producing systems.

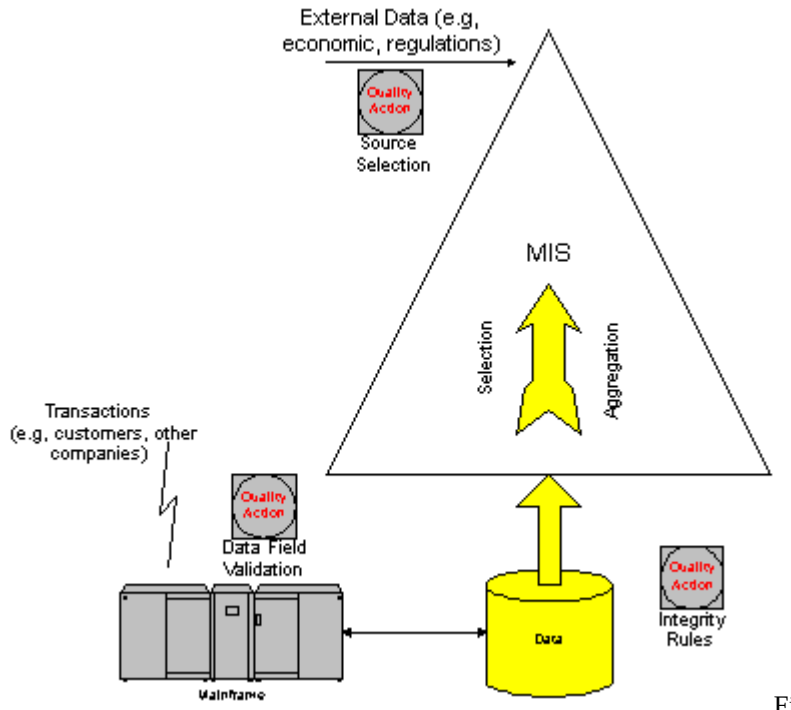

Figure 3: Traditional Information Flows

The proliferation of personal computing and Internet usage have changed the information flows, increasing the "unmanaged" flows dramatically, as illustrated in Figure 4. While information quality procedures improved for the transaction generated data (primarily due to increased use of electronic capturing and exchange of data, and integrity rule support by DBMS technology), serious quality problems arose for ad-hoc information generated from various sources, such as external Web sites and individual users' personal computing applications.

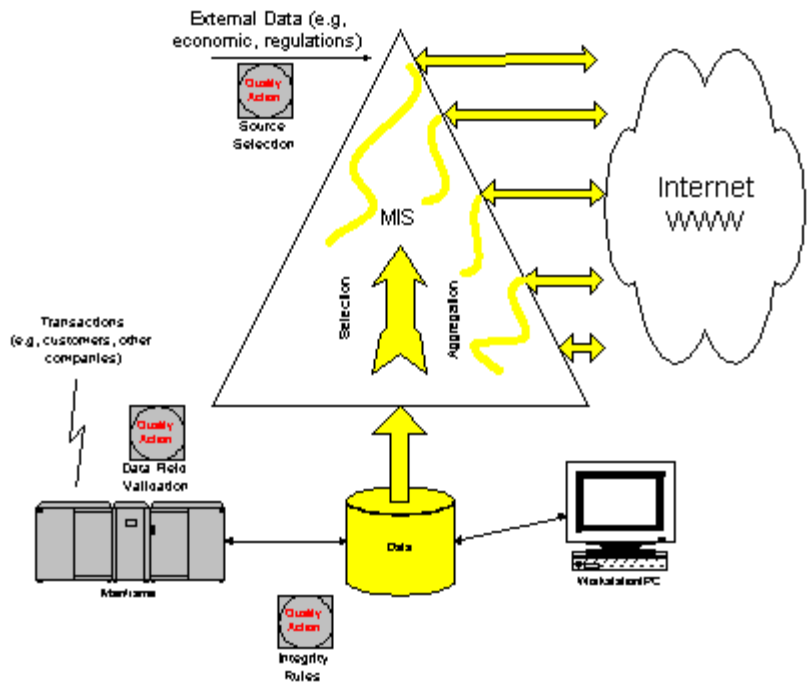

Figure 4: Today's Information Flows

It is evident that there is an exponential growth in external information flow (bi-directional, into the company and out of it)), business-to-business and business-to-consumer (government and non-profit organizations are considered "business" in this context). This information flow needs to be managed with respect to data quality.

One of the consequences of the emerging technologies is a mix of new data types that are not your traditional database types. They include HTML, SGML, EDI, MIME (and plain e-mail). These new types and their related technologies add significant complexity to the function of *Data Administration* and require changes to traditional definitions (at the database level) of data items quality.

1. The Screen is the Information

A new generation of technologies have an interesting impact on information quality (broadly defined). These emerging technologies, which can be termed "screen control" are targeted towards improved user's productivity by providing intelligent screen capture and manipulation tools (see Figure 5). An example is technology developed by AnySoft [1], that in general consists of three components:

- Intelligent contents capture from the screen (e.g., text, graphics)
- Manipulation and transformation of the captured contents (e.g., graphing, arithmetic)

- Sending the transformed contents (possibly "as-is") to selected destinations (e.g., files, databases, other applications)

Besides increased productivity, this technology is related to information quality in the following ways. It can significantly improve the presentation and usability dimensions, and it can be used to control the display of contents. On the other hand, like many other technologies, if not managed it can introduce quality problems with respect to source tagging or identification, since newly generated information can originate from screen displays rather than from applications. It turns out that this technology itself can provide the ability to trace and manage an important class of information flow -- the information that "passes through" the screen.

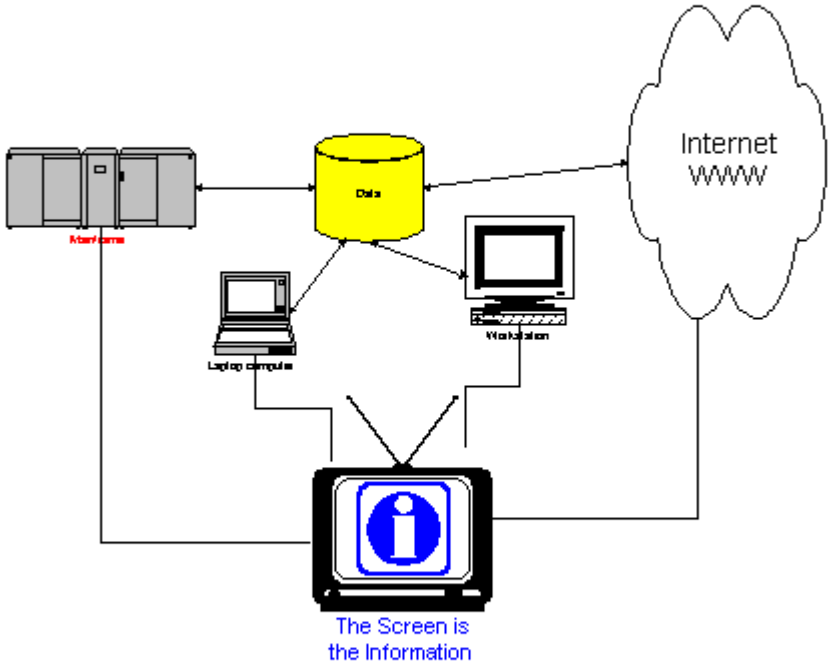

Figure 5: The "Screen is the Information" Concept

1. Conclusions

This paper emphasized the importance of identifying information flows and processes patterns in an organization - a prerequisite for managing information quality. It presented the notion of managed information flows and processes vs. unmanaged ones. Many of the proposals for information quality improvements assume the "managed" case. Unfortunately, a significant, and increasing, portion of the information flows and their processing is unmanaged. The Internet, WWW, distributed desktop computing, and screen capture and processing technologies, will make the situation worse, unless more of the flows become managed.

The good news is that the same technologies can be used to improve various data quality dimensions. WWW applications (both Internet and Intranet) reduce the amount of paper work and the number of steps involved in data capture; they also provide for direct input of the data by their "owners", e.g., customers change their address directly through the Web. Screen control technologies can be used in dual mode, on one hand providing the user with an ad-hoc mechanism to manipulate data in ways that pre-built systems would not allow, and on the other hand provide an administrative control that can be used to better control information quality.

More research is needed in several areas. Source tagging and process tracing are crucial; database schemes on their own can't be totally effective, and integration with other non-database technologies is essential; any resulting scheme should make sense with respect to organizational information quality processes. Allowing user choices in data retrieval (see [4] for example) can lead to higher-quality information in cases where the "common wisdom" calls for consistency resolution at the source (or as close to the source as possible). Measuring the quality of new data types (such as multi-media objects) is also an important area; obviously, a range check, such as in salary value validation, is not appropriate here. And finally, though not addressed explicitly in this paper, resolving data heterogeneity is a fundamental problem is data quality, and a more difficult one in the context of the WWW.

References

[1] AnySoft Office tools; AnySoft Inc., Cambridge, Massachusetts, 1996.

[2] Ballou, D. P. and K. G. Tayi. Methodology for Allocating Resources for Data Quality Enhancement. *Communications of the ACM*. *32*(3) 1989. pp. 320-329.

[3] Ballou, D. P., R. Y. Wang, H. Pazer and K. G. Tayi. Modeling Information Manufacturing Systems to Determine Information Product Quality *Management Science (forthcoming)*. 1996.

[4] Gal A., Etzion O., and Segev A. Representation of HighlyComplex Knowledge in a Database. *Journal of Intelligent Information Systems*, No. 3, pp. 185-203, 1994.

[5] Liepins, G. E. and V. R. R. Uppuluri, ed. *Data Quality Control: Theory and Pragmatics*. D. B. Owen. Vol. 112. 1990, Marcel Dekker, Inc.: New York. 360 pages.

[6] Redman T.C. Data Quality - Management and Technology, Bantam Books, 1992.

[7] Wand, Y. and R. Y. Wang. Anchoring Data Quality Dimensions in Ontological Foundations. *Forthcoming, Communications of the ACM*. 1995.

[8] Strong, D. M., Y. W. Lee and R. Y. Wang. Data Quality in Context. *Forthcoming in Communications of the ACM*. 1996. .

[9] Wang, R. Y. and H. B. Kon, Towards Total Data Quality Management (TDQM), *Information Technology in Action: Trends and Perspectives,* R. Y. Wang, Editor. Prentice Hall: Englewood Cliffs, NJ. 1993.

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

[11] Wang, R. Y., V. C. Storey and C. P. Firth. A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering*. *7*(4) 1995. pp. 623-640.