Grand Challenges in Data Integrity and Quality: Panel Discussion

Bhavani Thuraisingham¹

The MITRE Corporation, USA

** On Leave at the National Science Foundation, Arlington VA 22230, USA*

Abstract: This paper provides a summary of the panel discussions on data quality at the IFIP Integrity and Internal Control in Information Systems Conference held in Lausanne, Switzerland, November 2003.

Key words: Data Quality, Data Integrity, Data Security, Quality of Service, Semantic Data Quality, Data Quality Algebra

1. INTRODUCTION

Data integrity and quality have received a lot of attention recently. Data integrity and quality are about maintaining the correctness, accuracy and quality of the data. Especially in a heterogeneous and web-based environment, data may emanate from all kinds of sources. Therefore it is difficult to trust the accuracy and quality of the data. Data is however a critical resource in almost all organizations. It is critical that appropriate techniques are developed to ensure data integrity and quality so that data users can carry out their functions effectively.

¹ **DISCLAIMER:** The views and conclusions contained in this paper are those of the author and the author's understanding of the discussions at the conference. They do not reflect the policies and procedures of the National Science Foundation or of the MITRE Corporation.

While research in data integrity and data quality has progressed a great deal, there is still a lot to do. Although we now have data cleansing and data quality tools, we need a strong research program to advance the field to ensure the quality of the data in a web-based world. Because of the importance of the area we organized a panel at the IFIP Integrity and Internal Control in Information Systems Conference in Lausanne, Switzerland held in November 2003. The panelists were the following researchers (in alphabetical order):

Karl Aberer, EPFL, Lausanne, Switzerland
Yves Deswarte, LAAS-CNRS, Toulouse, France
Sushil Jajodia, George Mason University, USA
Leszek Lilien, Purdue University, USA
Indrakshi Ray, Colorado State University, USA
Monica Scannapieco, University of Rome, Italy

The panelists discussed many challenges. In section 2 we summarize the challenges. Some directions are given in Section 3.

2. PANEL DISCUSSION

The challenges discussed by the panelists are the following:

- Integrating security and integrity
- Quality of Service
- Secure high integrity transactions
- Semantic data quality
- Algebra for data quality
- Trust management and data quality

Below we will elaborate on the challenges for each of the topics listed above:

Integrating security and integrity: There have been many discussions over the years on integrating security and integrity. From a security point of view integrity is about unauthorized modifications of the data. Can we develop solutions to detect and prevent unauthorized modifications as well as malicious corruptions of data? Can we use data mining techniques to detect and possibly prevent faults? How can we develop software that is self-healing? How can we recover from the faults in a timely manner?

Quality of Service: Closely related to the first topic is developing quality of service primitives. We need flexible security and integrity policies. For example, how can we effectively integrate security, integrity, fault tolerance and real-time processing to achieve dependable computing? How can we make tradeoffs say between security, integrity, fault tolerance and real-time computing? For example, in some situations we may need 100% security while in some other cases it is critical that the transactions meet the timing constraints and therefore we may be willing to sacrifice some level of security. How can we specify the flexible policies and subsequently implement the policies in operational systems? How can we ensure that there is end-to-end security and survivability and ensure that systems that are composed are secure and survivable?

Secure High Integrity Transactions: There has been a lot of research on secure transactions and especially on multilevel secure transactions. The idea is for high-level transactions not to interfere with low level transactions and send covert messages. The challenge is to ensure that the integrity constraints and data quality constraints are satisfied by the transactions. There has been work on integrity constraint processing during transaction processing. However integrating security, integrity, data quality and transaction processing remains a challenge.

Algebra for Data Quality: Various algebras have been proposed for data manipulation including relational algebra. Can we extend these algebras to include data quality? One simple way to accomplish this is to have a data quality attribute associated with each attribute in a relation. Then we need to develop an algebra so that data quality values can be computed for the results. So the challenge is how can we develop algebra for manipulating data quality values as well as handle integrity constraints?

Semantic data quality: Database community has worked on semantics for a number of years. They have developed solutions to semantic heterogeneity, a problem that occurs when integrating heterogenous databases. The community is now moving in the direction of information integration where information from diverse disparate data sources has to be integrated on the web. Ontologies have been specified to aid information integration. The goal is to move toward the semantic web. Data quality plays a crucial role. In order to ensure that the data is accurate and consistent for the semantic web, we need to develop data quality techniques for the semantic web. We need to understand the nature of the data and consider the semantics when assigning data quality values. Essentially data quality has to be considered when discussing the semantics of the data. Ontologies have to specify data quality,

data integrity as well as data security features. In summary, ensuring data quality for secure information integration on the web remains a challenge.

Trust management: Security researchers are working on trust management issues. The challenge is how do you trust the data and information provided by the publisher? The publisher may have received the data from another source and that source may have received the data from yet another source? One's trust of the quality of the data depends on how much he or she trusts the sender. The sender in turn trusts the quality of data depending on who sends it to him or her. Can we develop a theory of trust? How can we carry out trust negotiations with respect to data quality? That is, can A and B negotiate with each other and develop trust mechanisms so that they can effectively share data and collaborate on tasks? How can we ensure trust and integrity in a peer-to-peer environment? There are many challenges in trust management that need to be addressed for data quality.

3. DIRECTIONS

The panel provided some good insights into directions and challenges in data quality and data integrity. While the panelists were from academia, there was also industry participation at the conference and we got some useful information on data quality for various industries. For many applications in the commercial world, including defense, intelligence, banking, financial, and medical areas, data quality is critical. How can we act on the data for critical applications in medicine, finance and defense if we are not certain of the quality? There was a general consensus that while everyone agrees that data quality and integrity are critical, the areas have not received a lot of attention in the research world.

The challenge is for the various research communities to actively pursue research and convince the research funding organizations that there is a clear need to start research in this area. It was noted that there are now many research funding initiatives in both data management and information security. The question is, for now at least should we focus on data quality and data integrity as part of the research programs in data management and information security? This way the area will get some focus. Then as we make more progress and uncover more challenges, we could devote more resources in this area.

The group also felt that the panel has raised many interesting ideas and discussed some key challenges. We should incorporate some of the ideas presented in the call for papers for next year's conference so that we can keep the research going in this very important field.