INFORMATION QUALITY ANALYSIS

Bahador Ghahramani, Ph.D., P.E., CPE

206 Engineering Management
School of Engineering
University of Missouri-Rolla
Rolla, Missouri 65409-0370 (USA)
E-mail:ghahrama@shuttle.cc.umr.edu

INTRODUCTION

The rapid proliferation of state-of-the-art electronically stored data, the information super- highway, advanced information systems (IS) that input and generate data, and modern Systems Engineering (SE) improvements have increased the importance of information quality in the Twenty-First Century communications age. It is a fact that, despite improvements in information efficiency, speed, accuracy, verification, and validation processes, output quality in IS is usually poor and often unacceptable. In most cases, when information quality problems are considered and data accuracy issues are addressed, resources are allocated toward correcting the faulty data rather than improving the process.

Information quality analysis makes a comparison between the data provided and "finished" products or services at the end of a production process. The distinction will be made that, in real-time environments, data are continually updated and validated. The accuracy of one piece of information, therefore, may not depend on correctness of another; and, correction of one faulty value may not prevent the occurrence of others at a later point in time. However, validity of the outputs as a whole depends on accuracy of each piece of data. In IS environments, a correct value may be accurate for a limited duration of time, becoming faulty or outdated at a later time.

This paper will also address the following fundamentals of data quality and future evolutionary initiatives in modern IS.

- Processes that deliver data and the accuracy of the data become more significant. This includes on-line services such as automated processes that enhance research and development efforts.

- Sophisticated business decisions and complex managerial planning significantly rely on information processes such as operations, finance, sales, production, and the quality of the results.

- As the number of information users increases, along with speed and volume of data, system environments will become more complex and sophisticated.

BACKGROUND

Corporate decisions are increasingly based on data stored in databases. High level term plans, mergers, reorganizations, and vital initiatives are decided through aggregation and disaggregation of vital information. In corporate America, databases are regularly used to generate reports, and to make numerous vital decisions. How accurate, pertinent and valid are these millions of reports generated and corresponding decisions? As accurate, pertinent, and valid as the data inputted in the databases. If database information lacks accuracy and validity, it can cause severe problems. These problems may result in poor allocation of resources, inefficiencies, and ineffectiveness which can cause the system to become irrelevant.

Systems Engineers estimate that error rates of 15% to 50% are common in databases in a wide variety of IS applications. However, these estimates do not reveal the true extent of the problems and their associated costs. These estimates only reveal the dimension of data quality - they do not address the loss of opportunities and their adverse human impacts. They do not indicate whether data records are current, complete, or consistent across the board.

The SE approach consists of a systematic data collection procedure, accuracy check, and process validation. This approach also improves understanding of the subprocesses and provides an objective realization of decision making activities. The SE approach covers scientific tools such as Operations Research (OR) techniques; and Statistical Quality Control (SQC) methods. SE evaluation of the results are paramount in the light of sound scientific tools, experience, objective judgment, expertise, and clear comprehension of the process. This SE approach benefits the IS process through analysis of:

- Variations in the process, at the beginning stages, in time to correct a problem before it adversely impacts the system;

- Process flexibility to meet system's requirements and objectives; and

- Resources and efforts to set improvement goals and to initiate effective solutions.

IS data quality analysis requires a through understanding of the data input process, database capabilities, output accuracy, system's integrity, and process productivity. Figure 1 presents the relationship between data inputs, outputs, and the database. As this figure indicates,

the most important factor in data accuracy is that the information must be “clean and accurate” before it is transmitted into the database. In most applications it is costly, inefficient, and impossible to test the accuracy of every piece of data. At the same time, a poor or unknown quality of information may cause great harm to the decision making process. To improve the quality of the data, a SE plan must be initiated and activated that maintains the highest level of standards and process integrity.

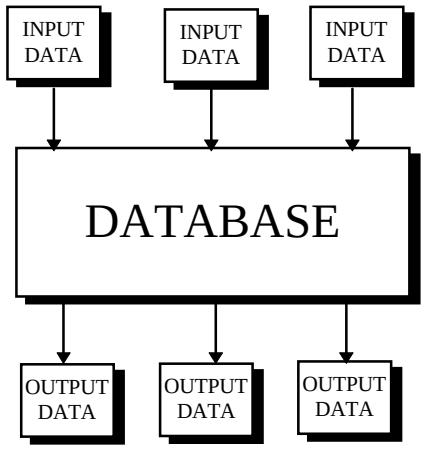

FIGURE 1, RELATIONSHIP BETWEEN DATA INPUTS, OUTPUTS, AND THE DATABASE.

PROCESS ANALYSIS

The SE approach entails tracking data as it is processed through the system in order to identify the occurrence of distortions. Since it is inefficient, time consuming, costly, and ineffective to track and evaluate every bit of data as it is processed, scientific tools and SQC are used to take random samples of input data. If correctly taken, samples of the data can discover error patterns that, when considered, will lead to improvements in the accuracy of the process. A pattern of errors in the sample data indicates the root causes and helps to identify probable solutions. It is then possible to focus on the causes of the problems and eliminate them before they reach the database.

Tracking data through the process requires two initiatives: (1) selecting the pieces of targeted data and (2) monitoring their movements through each subprocess of the system. Figure 2 shows data movements through various subprocesses and entering the database.

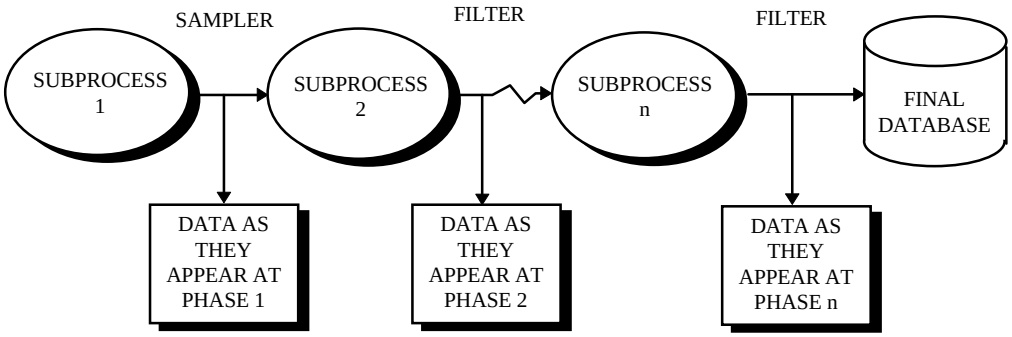

FIGURE 2, DATA MOVEMENT THROUGH THE PROCESS.

To accurately follow information through the process, it is essential to implement an effective data tracking system (DTS). As data flows through the system, the DTS documents the movements of data through the process, records their values at each subprocess, and tests their accuracy before reaching the database. Comparison of the records at various subprocess points makes it possible to identify data irregularities, inaccuracies, and variations. Figure 3 addresses the DTS initiative - discussing data movements through the system starting from the beginning phase (phase 1) to the final phase (phase n). DTS performs the following efforts:

- Describes the data flow process - identifies key subprocesses and their functions;
- Initiates the data tracking plan - defines the DTP phase requirements and measures;
- Analyzes and validates samples of collected data - evaluates conformance to the requirements and standards;
- Evaluates the process - identifies and prioritizes problem areas;
- Implements scientific tools - random selection of records at various points and performs data accuracy test at different phases; and
- Improves data accuracy and validity - applies scientific and SQC tools to continuously improve the process.

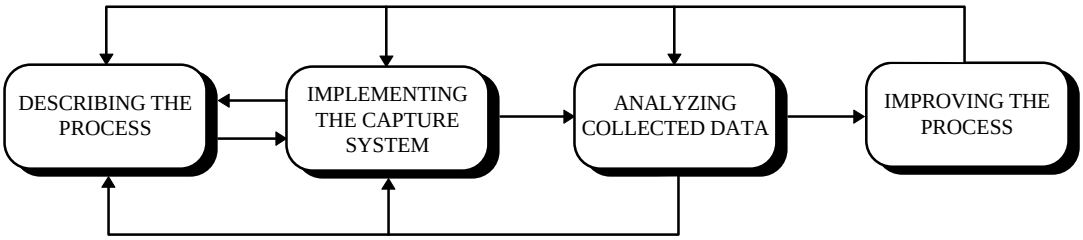

FIGURE 3, DATA TRACKING SYSTEM.

Figure 4 is a flowchart of the DTS analysis. This figure maps the data from user input to user database as it goes through subprocesses, samplers, and filters. As this figure illustrates, user inputs pass through various points and ultimately reach the database.

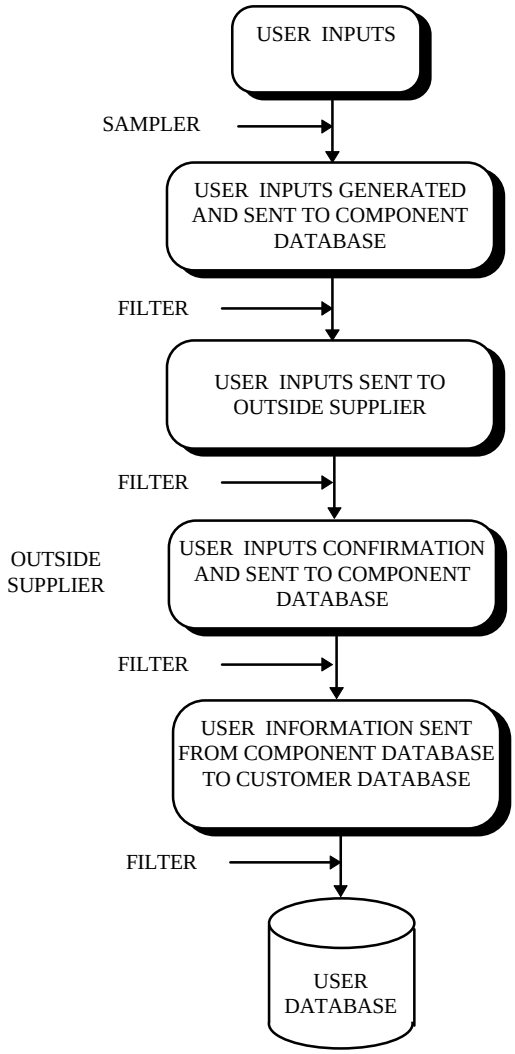

FIGURE 4, DTS ANALYSIS

Describing the flow process requires identification of the data origin and the subprocesses it follows to reach the database. It entails the data gathering points at every subprocess and phase. The DTS also describes the data assembly process and the sequence of phases the information follows through to generate the final outputs. Various scientific tools are used to describe the data flow process: the Functions of Information Processing (FIP) method, flowcharting, data-flow diagrams, and other block diagrams. All of these tools require a clear understanding of the data assembly process and realization of where various types of data are originated. Identification of the origin of data, and how it was started, helps to better address the problems. To determine the origin of the data, it is essential to identify and analyze:

- Type of data - numeric, alphanumeric, or binary;
- Length of data - in characters or bytes;

- Purpose and use of the data - why it was produced and for what reasons;
- Methods selected to gather the data - electronics, hard copy, analog, and digital; and
- Sequence of operations performed - coding and formatting.

DTS identifies origins of the data, determines the exact characteristics of the data assembly process, and indicates hardware and software tools used. To gather information throughout a system, there is a need to follow a SE approach to data assembly process. This SE approach consists of:

- Random selection of records entering at subprocess points;
- Capture data from the selected records at designated subprocess points; and
- Identify the system samplers, filters, and database.

Figure 5 is similar to the previous illustration, it studies the physical location of the database throughout the system.

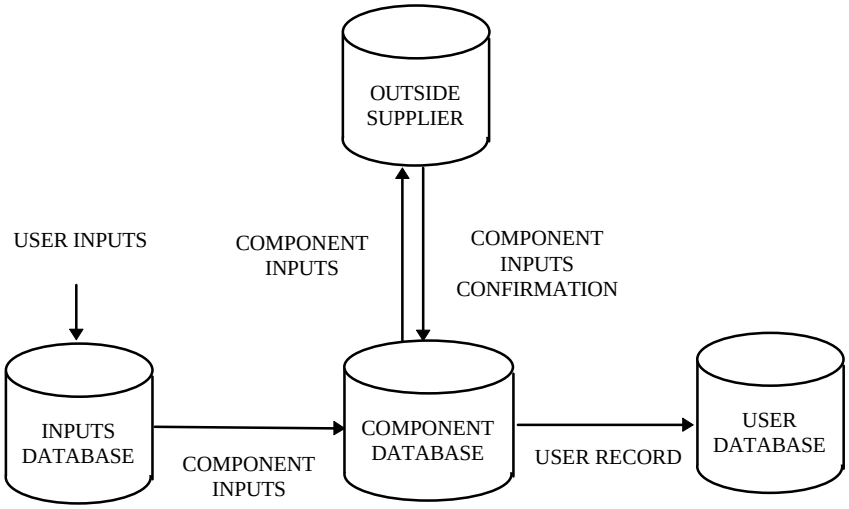

FIGURE 5, PHYSICAL FLOW CHART OF DTS.

SAMPLER AND FILTER

A sampler selects a sample of records as they enter the designated subprocess points. The sampler is part of the data assembly process; it is used for randomly selecting data from records at different time intervals to create samples for further observations. The sampler captures data

at different time intervals from selected records, identifies, evaluates, and records the time and corresponding subprocess. This DTS can be automated using a sampler program that can perform the same functions more efficiently and accurately. Figure 6 is a presentation of an automated DTS.

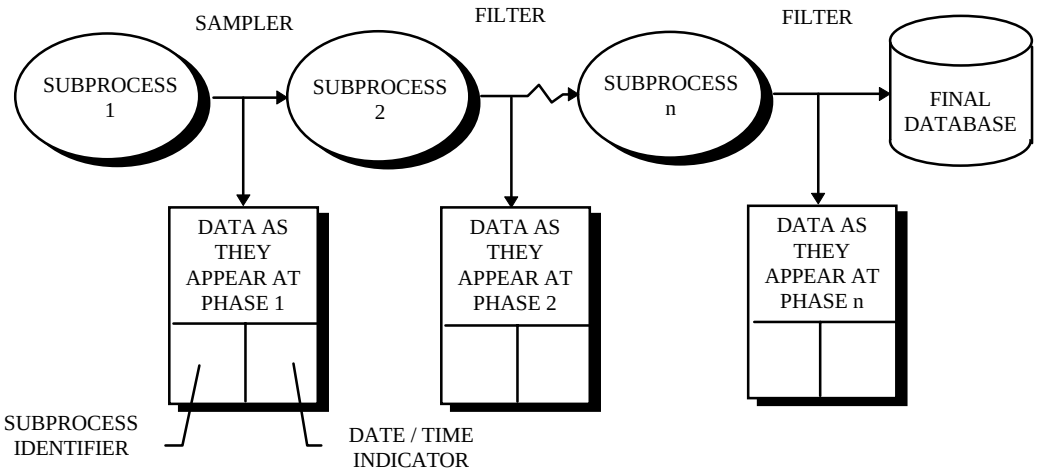

FIGURE 6, AUTOMATED DTS.

A filter is a part of the data assembly process that distinguishes, separates, and pulls out the selected records being tracked at the sampling points. It captures the data on the records together with some identifier such as date and time. For the filter to perform its tasks, the sampling points must be at the output of the previous subprocesses. The filter functions are automated using a filter software program. Using filter software increases the efficiency and accuracy of the process.

Sampling of data depends on the process of its arrival at the designated sampling points. There are two primary types of data arrivals: continuous and batch. Continuous data arrivals occur when records arrive at a sampling point randomly, over a period of time, and without any group characteristics. In this type of arrival, records are not in groups and arrivals are not cyclical.

Batch arrivals are when records arrive in distinctive groups spaced in a period of time. These type of records share some common characteristics and their arrivals are predictable. Batch arrivals are easier to organize by: place of origin, type, characteristics, time, and date of the arrivals. It is always best to select batches of more than 50 records for a sample so that SQC tools can be applied.

To improve the accuracy of the continuous data, there is a need to develop a continuous sampler at the sampling points. Most continuous samplers are effective if they are programmed to operate for a long period of time to gather 50 to 100 randomly selected records each time a sample is taken. This can be effectively done through selecting:

- An appropriate time interval to sample more than 50 records;
- Determine the average number of records arriving in the selected interval; and
- Find F, the fraction of arriving records, that is required to sample 50 to 100 records per period. F is computed based on M, the average number of arrivals.

$$F = \frac{67}{M}$$

Where 67 (2/3 of 100) provides a high probability value that the number of records sampled in the interval will fall between 50 to 100.

Integrating the capture system consists of assembling a host of systems with varying combinations of software and hardware based on their degree of automation. This degree of automation distinguishes the characteristics of the capture system. Figure 7 is a presentation of a systems' categories at different degree of automation.

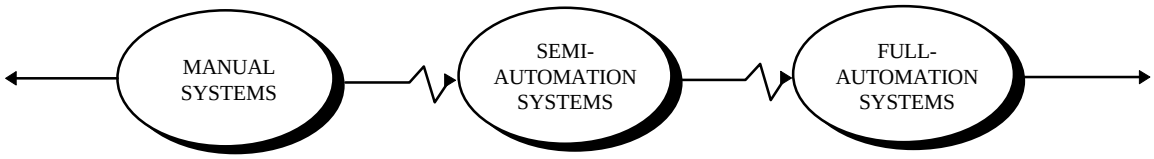

FIGURE 7, SYSTEMS AUTOMATION CATEGORIES.

CONCLUSIONS

This paper scientifically analyzes information quality, its utility, and addresses a SE view of the overall process and identifies opportunities for continuous improvements. The IQA presented here uses SQC and OR techniques to increase quality of an information system and enhance its added value to the users and customers. The IQA is based on a SE approach that tracks data as it is processed through the IS and identifies occurrence of variations, errors, distortions, and bottlenecks. It provides the Systems Engineers with an automated DTS that is able to take random data samples at designated points within the system to discover error patterns that can lead to degradation in the accuracy of the process. A SE analysis is able to identify root causes of errors in the samples to develop a formula to eliminate the problems. DTS continuously monitors, evaluates and documents the data flow at random time intervals through the process, records their values at each subprocess, and tests their accuracy before reaching the database. The DTS also identifies origin of the data, determines the exact characteristics of the data assembly process, indicates hardware and software used, and the sequence of phases the information follows to generate the results.

The presented IQA addresses a physical flowchart of DTS, and the importance of modern samplers and filters, in an automated and state-of-the-art IQA process. It highlights the distinction between the two primary data arrivals, continuous and batch, and their importance to the IQA process. Figure 8 is an overall view of the IQA presented in this paper. As this overview shows, the process follows an SE approach that starts with regrouping records to analyze the overall process performance and completes the process by making a detailed analysis of all activities.

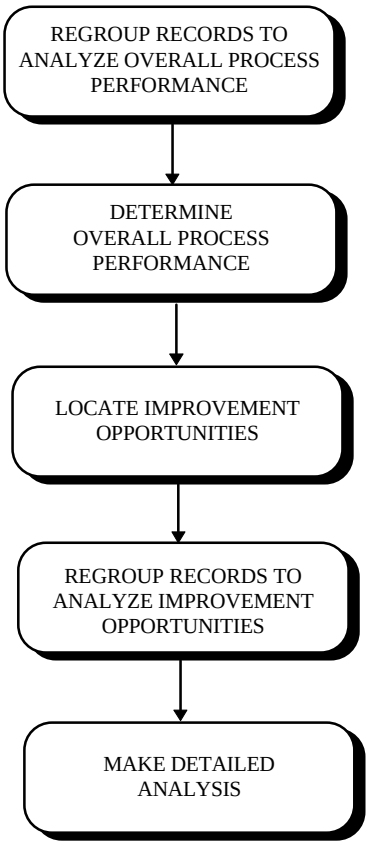

FIGURE 8, AN OVERALL SE APPROACH TO IQA PROCESS.

ACKNOWLEDGMENTS

The author wishes to express appreciation for the support of his colleagues in the School of Engineering at University of Missouri-Rolla, and to graduate students Phillip B. Swan and Richard Chi-chung for their inputs. Also, his sincere gratitude is given to Systems Engineers in Bell Laboratories and IBM Watson Research Center for their reviews and recommendations.

REFERENCES

[1] Widmann, E.R. "Capability Assessment Model for Systems Engineering", Proceedings of the Third Annual International Symposium of the National Council on Systems Engineering, 1993.

[2] Mackey, Dr. William "Conducting a Systems Engineering Process Assessment", Proceedings of the Fifth Annual International Symposium of the National Council on Systems Engineering, Volume I, 1995.

[3] Widmann, E.R., Andrews, B.A., D.L. Brenchley, and J. Worl "A Second Systems Engineering Process Assessment at the Westinghouse Hanford Company", Proceedings of the Sixth Annual International Symposium of the National Council on Systems Engineering, Volume I, 1996.

BIOGRAPHY

Dr. Bahador Ghahramani is an Associate Professor of Engineering Management in the School of Engineering at University of Missouri-Rolla (UMR). Prior to joining UMR he was a Distinguished Member of Technical Staff (DMTS) in AT&T-Bell Laboratories. His work experience covers several years of academics, industry, and consulting.

Dr. Ghahramani has presented and published numerous papers and is an active participant and officer of various national and international organizations and honor societies. He holds a patent, "Eye Depth Testing Apparatus", has filed for two Bell Laboratories patents "A Method for Measuring the Usability of a System" and "A Method for Measuring the Usability of a System and for Task Analysis and Re-Engineering".

Dr. Ghahramani received his Ph.D. in Industrial Engineering from Louisiana Tech University; MBA from Louisiana State University; MS in Industrial Engineering from Texas Tech University; MS in Applied Mathematics from Southern University; and BS in Industrial Engineering from Oklahoma State University.