Quality Of The Data Obtained In The Acquisition Process

Adrian Grigorovici, Ion Ivan, Gheorghe Noşca
Department Computer Science in Economics
Academy of Economic Studies, Bucharest, Romania
Piata Romana 6, Bucharest, Romania
ionivan@ase.ro

***Abstract*:** Defining the data quality concept. Building a data acquisition system and identifying acquisition processes. Developing the data quality control system. Establishing the risks and errors that affect the acquired data quality.
Keywords: data quality, data acquisition, QoS – quality of service

1. Introduction

A Data Acquisition System, DAS, measures and records some specific parameters in order to analyze the storage data and to improve the measured characteristics. DAS has both a hardware component, represented by the data acquisition devices, computer systems and the connection between them, and a software application for acquisition, storage, and analysis of data.

Data and information have an increasingly importance into a knowledge based society. A data error, in nuclear field or in air flights traffic control, for example, could lead to a disaster with catastrophic consequences. Solutions for the errors' elimination are searched for the entire data usage process period. One of them is the error generator factors' elimination in data acquisition process. In order to analyze and eliminate the errors, it is necessary a data quality and influence factors' analysis. The acquired data analysis leads to software improve used in the application.

2. Data quality features

Quality is a fuzzy concept, complex, derived from the aggregation of characteristics with different measurement units. Data quality is an issue whose importance is growing for business strategies.
To analyze the data quality, it is necessary to establish a set of features that defines various data aspects. In practical activities, it has been identified over 150 data quality features. In order to a better quality characteristic analysis, researchers in the field grouped the data quality features into categories and sizes. The size is a features quality set to which users react in a consistent manner. The most important data quality dimensions are shown in the Table 1.1.

Tabel 1.1 Categories and sizes of the data quality

Data quality categories	Data quality dimensions
intrinsic	accuracy, objectivity, credibility, reputation
accessible	accessibility, access security
contextual	relevance, added value, opportunity, completeness, volume
representation	interpretability, easy understandable, concise representation,

	consistent representation

In practice, it is found that the data, in the acquisition process, are influenced by the following characteristics:
- *accessibility* assumes that the data are easily available in a short term, search and retrieval algorithms were successfully completed using various criteria, including the keys' search incomplete definition;
- *access security* refers to the degree to which data access is restricted and, thus, protected; security raises barriers in the way to accessibility;
- *correctness* is the degree to which the data are certified as being without error;
- *accuracy* of data refers to the approximation of the value *v* with a value *v'* in the field of characteristics (*v*' is considered correct for the entity *e* and feature *a*). If the value *v* coincides with the *v*', the data is correct;
- *objectivity* is the degree to which data are undamaged and impartial;
- *credibility* lies in the degree to which data are accepted or regarded as true and real;
- *actuality* refers to the degree to which a data is updated;
- *consistency* means that two or more data are not in conflict with each other;
- *reputation* is the degree to which data are rated in terms of their source or content;
- *relevance* refers that the provided data are necessary to the application which have been supplied;
- *added value* is the degree to which the data bring benefits and provide advantages through their use;
- *opportunity* lies in the degree to which the data age is appropriate for the performed activity;
- *volume* is the degree to which the amount or volume of data is appropriate;
- *completeness* refers to the degree to which data values are present in a collection of data; [IVAN06]
- *complexity* of data [IVAN99a] has a high degree of influence on data quality. To measure the complexity it is necessary to consider the diversity of data types, the number of impressions and the links between data;
- *orthogonality* of data shows the measure to which an element a_{ij} in a community is different from another element a_{ik}, or differs from all other elements forming A_i collectivity.

The developing of applications in the Internet resulted in the emergence of new users requirements. The data quality from a Web site is supplemented by:
- *structurability* refers to the ability of a web application to provide Internet users an intuitive mechanism, in relation to his browsing experience, which he found the requested information. The information finding duration is inversely proportional to the structurability degree;
- *authenticity* assumes the user security in using available application data;
- *level of noise* refers, in the Internet applications, to the existence of sounds, images, texts undesirable by the user.

Basic categories of the data quality characteristics and measurements on them are grouped in inherent and pragmatic.

Data quality inherent characteristics are: definition compliance, values' completeness, validity or compliance business rule, source accuracy, reality accuracy, precision, non-duplication, redundant or distributed data equivalence, redundant or distributed data competition.

Data quality pragmatic characteristics refer to more efficiency that the data get in activities' deployment by those in the informatic field.

3. Data acquisition devices

a) *Keyboard* is a hardware computer component which allows to an operator to introduce, by keypressing, from th console

figures, letters, miscellaneous control and command characters, encrypted as scanning code. The different keypresse (keystroke) combinations represents computer acquired data and commands addressed to the computer. The standard keyboard is called QWERTY.

It exists virtual keyboards also, which project an image, in natural size of the keybord, on a specific surface. The sensors on that surface transmit to the computer which key was pressed.

b) *Mouse* is a hardware computer component used by an operator to send it commands. The mouse allows the cursor positioning in the monitor pixel matrix and the displayed options selection.

The mouse devices could be: mechanical, optical, infrared optical, laser, inertial, 3D or tactile mice. Also, the mouses could be with 1, 2, 3, 5 or more buttons.

c) *Touchscreen* is a display which detects the presence and location of a touch on a display surface. When we say touchsceen we refer to touch or contact, by a finger, hand or passive object like stylus, to the display of the device.

Touchscreen has 2 important features:

- it enables you to interact with what is displayed directly on the screen, where it is displayed, rather than indirectly with a mouse or touchpad. Secondly, it lets one do so without requiring any intermediate device, again, such as a stylus that needs to be held in the hand. Such displays can be attached to computers or, as terminals, to networks. They also play a prominent role in the design of digital appliances such as the personal digital assistant (PDA), satellite navigation devices and mobile phone.

- Allow direct user interaction through the intermediary of the stylus. There are several types of technologies stylus: panel touchscreen rezistiv; sensitive to surface acoustic wave, surface acoustic wave, SAW; capacitive touchscreen panel, the panel touchscreen with infrared sensors IR military applications, image sensors, optical imaging, the signal dispersants; recognition with acoustic impulse; interruption with total internal reflection.

d) *Video camera* is a device used for the acquisition of images in motion. 2 are the most important features of cameras: it can transmit images directly in motion and can store images recorded for archiving and subsequent processing. There are several types of professional video cameras such as camcorders, which includes a video camera and a VCR, Video Cassette Recorder, or other recording device, television cameras closed-circuit, CCTV, Closed-Circuit Television Camera, digital cameras, Digital Camera System for the cameras, such as those on board satellites and spacecraft sondelor or used in research in the field of artificial intelligence and Robotics.

e) *Scanner* is a device that transfers images graphics, text written by hand, printed text or objects in digital format. Most used scanners are: the image scanner, converted into a digital image dimensional, 3D scanner, converted into a digital three-dimensional image of a real object, film scanner, or turns negative or positiv film in a digital image.

The term was extended scanner on any device or software that performs a search, such as radio scanner, the frequency of searches for the reception of a radio program, the scanner tape, searching for breaks between records, rotating radar antenna, vulnerability scanner, a software looking for weaknesses of a system, analizorul lexical, software analysis of the text, port scanner, looking for open ports of a computing system, LIDAR scanner for flight 'time-of-flight and triangulation 3D laser scanner, which uses scanners active laser light to investigate buildings, geological formations in order to achieve a 3D model etc..

f) *Tape card reader* is a device that reads data stored on a tape with magnetic material located on the surface of the card, by altering the magnetism of small

particles of steel. Magnetic tape is read by card reader with the tape through physical contact. Cards with magnetic tape used in credit cards, identity cards, transportation tickets, etc.

g) *USB* is a data storage device that uses standard USB connection, bus and a USB connector, Universal Serial Bus. The data on this device can be transferred through orders given by the operating system to computer attached to the USB device.

h) *CD Unit*, Compact Disc, done reading optical data from the CD. CD is a device used to store data in digital format. CD was originally created for information storage and audio to replace the floppy drive, but has lower environmental performance storage on the hard disk.

CD has more constructive options: CD-ROM, CD Read Only Memory, used only for reading data already entered, CD-R, CD Read, used for registration only once and then only for reading data, CD-RW, CD Read / Write, CD reads and writes to multiple data, SACD, Super Audio CD, VCD, Video Compact Disc, SVCD, Super Video Compact Disc, Photo CD, Picture CD, etc.

i) *DVD Unit*, Digital Video Disc, provides reading optical data from the DVD. DVD is an optical disc, which succeeds CD, with a storage capacity greater, at 4.7 GB DVD vs. 0.7 GB CD, and a different way to write information on the disk ..

DVD has several constructive variants: DVD-ROM, DVD Read Only Memory, used only for reading data already entered, DVD-R and DVD + R, DVD Read, used for registration only once and then only for reading data, DVD RW, DVD Read / Write, DVD reads and writes to multiple data etc..

j) *Microphone* is a signal translator that converts sound pressure air into electrical signals. There are several types of microphones: capacitive microphone, condenser / capacitor microphone, turns the air pressure caused by sound change in the ability of a condenser, and then the electric signal, dynamic microphone, dynamic microphone, air pressure caused by sound is converted into mechanical vibration a skirt microphone which turns, electromagnetic induction, the electric signal; piezoelectric microphone, piezoelectric microphone, air pressure caused by sound is converted into mechanical vibration of the piezoelectric crystal microphone, which is converted into electric signal; microphone with laser, laser microphone, a laser beam falls on an area subject to vibration of air caused by sound, a receiver takes power variations of the laser beam bounced corresponding vibration area, which are then converted into an electric signal;

Microphone can be omnidirectional, unidirecționale or shotgun, the most professional microphones unidirecționale. The sounds are taken with the microphone, converted into electrical signals and acquired data storage.

k) *Bar code scanner*, is an optical device that reads the data encoded in barcodes. Barcodes are found in two representations: linear bar code, 1D or 1-dimensional, consisting of parallel lines of different thicknesses and the spaces between them, and matrix codes, matrix codes, 2D, 2-dimensional or bidimensionale, formats of points, hexagoane and other geometric figures, with greater capabilities data representation.

l) *Traditional photo camera* realise photos representing data stored on photographic film support of 35 mm. The photos, stored in a memory device, are made with digital photo camera. Digital camera, Compact Digital Camera, may store a large cantity of images in memory, make photos, record videos with sound, clear images for free storage and display images on the screen immediately after being recorded.

Digital cameras can be embedded in mobile phones, PDAs, Personal Digital Assistant, vehicles, satellites, spacecraft, etc..

m) *Sensor* is a device that measures the amount of physical size and converted into

a size that can be measured with a measuring instrument. Sensors are included in the transducers and can change a form of energy into another. Sensors can be classified by type of energy that it detects:

- Thermal sensors are to determine the variation in temperature: thermometer, termocuplu, thermistors, resistance temperature sensitive, thermostat, the bimetal thermometer, and to determine the variation of heat: calorimeter, sensor flow of heat, bolometer, which measured the incidence of radiation energy electromagnetic;
- Electromagnetic sensors are to determine the variation in electrical resistance: ohmmeter, multimeter; to determine the variation of power: galvanometer, ammeter; to determine the variation of voltage electricity: electroscope, voltmeter, to determine the variation of electrical power: wattmeter; to determine the variation of Magnetism: magnetic compass, compass Flux magnetometer, Hall effect device, to detect metal, radar, Radio **D**etection **a**nd **R**anging;
- Mechanical sensors to determine the variation of pressure: altimeter, barometer, barograf, air speed indicator, variometru, speed indicator rising, pressure gauge, to detect gas and liquid flow: the flow sensor, anemometru for measuring wind speed, fluxmetru, gazmetru, debitmetru, watermeter; to determine the variation of viscosity and density of gas and liquids: hydrometer, mechanical sensors: acceleration sensor, Sensor position switch, tube-shaped oscillating U, measuring strains, to determine the moisture variation: hygrometer;
- Chemical sensors are to determine the chemical variation of proportion: oxygen sensor, ion selective electrode, pH glass electrode, phmetru, detection of carbon mono oxide, carbon redox; to determine the variation of smell: Sensor-tin oxide gas, sensor of quartz crystal for microbalance;
- Radiation optic sensors are used to detect the flight of light, to determine the variation of light, fotodetectors: fotocell, photodiode, fototransistor, image sensor, to determine the variation of infrared light, proximity sensors, laser sensor to scan ; Fiber optic sensor; interferometric sensor, which diagnoses by studying the properties of wave pattern of interference created by superposition; scientilometer sensor, measuring atmospheric optical disturbances due to variations in temperature, humidity and pressure;
- Ionizing radiation sensors are to detect radiation: Geiger counter, dosimetru, a counter with scientilație, measuring ionizing radiation; sensor to detect subatomic particles: particle detector, scientilator, a substance that absorbs high-energy electromagnetic radiation or ionized by charged particles and, in response, causing fluorescence photon, releasing energy absorbed previously, the particle detector radiation ionizing; detector electrically charged particles Detector electrically charged particles; room with bubbles, a vessel filled with liquid transparent overheat, used to detect electrically charged particles that move through it;
- Acoustic sensors can be acoustic: SONAR (**So**und **N**avigation **A**nd **R**anging), and to determine the variation of sound: microphone, hydrophone, seismometer;
- Other types of sensors are to determine the variation of movement: speedometer, speedometer, radar gun, radar gun, odometru or milometru; to determine the orientation: gyroscope, with the ring laser gyroscope, the attitude indicator, attitude indicator, to determine the variation of distance: magnetostriction, a property of ferromagnetic materials causing change their shape when placed in a magnetic field, etc.

4. Risks in the data purchase. Errors' emergence. Criteria errors' classification

The error is the difference between the actual and the codified value of a specific characteristic of a given entity. The real value involves the existence of a measurable objective reality.

Risks in data acquisition are related to the data collection way. The procedures for data collection [IVAN96b] are:

- *direct collection*, in which the data are collected, either automatically through the machinery or directly by the human operator. So, it is realized the link between working and data processing systems. It has a higher degree of errors' induction;
- *indirect collection*, which consists in data marking for manually reading via keyboard or specialized sensors.

It is recommended computer and other instruments use, in the process of collecting and monitoring data, which achieve:

- errors' elimination, caused by human errors, including: measuring devices reading errors, data transcription errors in the primary documents, figures and letters reading errors. Data entry is an important activity, with a major influence in ensuring data quality;
- human effort focusing on the data analysis and interpretation, which lead to an increased efficiency of the human resources' use.

Errors in the production data are caused mostly by the defective design and inappropriate processes' management of the information production that generates information low quality as a raw material. Some errors' causes in the data production are presented in Table 3.1.

Table. 3.1 Errors' causes in data production

Deficiency	**Causes**	**Affected dimensions**	**Organizational effects**	**Fix**
different values for the same data	multiple sources	consistency, reliability	legal and financial problems	development of common definitions and consistent procedures
data loss	systemic errors in data production	fairness, completeness relevance	lost or distorted information	processes' statistical control, processes' improvement, behavior control and proper incentives
difficult data access in reasonable time	large amount of stored information	concise representation, opportunity, added value, accessibility	excessively high duration to extract the information	rewriting using the graphical user interface and customer systems' power
definitions, formats and inconsistent values	heterogeneous distributed systems	consistent representation, opportunity, added value	inconsistent data that are difficult to access and aggregate	data warehouse

useful data change	users' tasks and in the organizational environment changes	relevance, added value, completeness		anticipation, changes in the users' tasks and processes and systems review before the inadequacy generate crisis
limited data access	insufficient calculation ressources	accessibility, added value		policies development of policies modernization, so that consumers know when to expect more resources

Errors' causes in the data production are the direct result of how the data and processes' broadcasters work. Poor planning and inadequate management of the data production processes generate a poor data quality.

The errors that occur in the processes of collection and data entry are: [IVAN96]

Supplementing errors of the primary documents consist of slipshod transcripts, incomplete transcripts of the data, reversed figures, letters and names, data replacement with other data, failure units.

Key errors encountered are letters or numbers merging and erroneous typing. These errors are caused by *inattention* and *understanding mistakes*.

Time errors are caused by delays occurred between the data collection time and the system correction execution time. Delays that generate errors occur between the following moments:

- collection - centralization;
- centralization - introduction to the processing;
- input - output results;
- show results - information decision factor;
- decision - decision practical implementation.

Errors that occur during the data lifecycle are caused by the fact that the correct data value may become invalid with time.

Implemented measures, leading to prevent and eliminate data errors, are shown in Table 3.2.

Table 3.2. Measures to prevent and eliminate data errors [IVAN96]

During collection	During the introduction in the processing system	After introducing
- primary documents design - completed documents verification - staff specialization and training	- taking the correct data from the document - intermediaries'	- validation program

- conditions' creation for correct completion - ensuring the optimum volume of operations and the number of people participating in the collection and processing	elimination - double introduction - scanning	

Storing the same information in several places is an error generating source because it is difficult to ensure a consistent update of all copies.

5. Ways to increase the data quality

Intermediary rings' elimination in the data introducing chain in the processing systems means the elimination of the potential sources of input errors.

In order to achieve the data quality improvement, it is necessary to separate the various issues associated with data such as the intrinsic properties and the data collection and delivery systems.

In the data acquisition process, the ways to increase the data quality are applying to the following factors:

- human operators through
  - staff appropriate selection;
  - employees' training;
  - creating appropriate working conditions for the data introducing operators so that their attention should be focused on the activities only;
  - activity operators' verification;
  - motivating salaries and wages' correlation with the operator errors' number.

Each data operator activity is recorded and there are counting the entering data errors' types separately. The operators' income are correlated with the entered data quality.

In order to achieve a entry data quality high level, the operators with 0 errors are stimulated. It is shown also:

- the most common errors;
- causes that provoke these errors.

The data quality analysis is permanently.

- equipment by:
  - devices' choice and data input equipment with the quality performance according to the needs;
  - equipments' installation according to the regulations and providing an equipments' quality service;
  - equipment calibration;
  - equipments' operational checking in normal operating regime.

By using the scanning, the document is treated as an image. The computer "learns" to recognize letters, figures and other characters.

6. Conclusions

The application complexity is given by a factors' combination whose influence is determined by the features' quality levels. In many cases, it has been studied in correlation with the reliability, maintenance, and stability.

A particular importance in this analysis is the quality of the data. Because of the costs they generate in an organization and the fact that data generating additional costs necalitative large processing, data quality becomes a priority of any successful management. The vast majority of these attributes of quality are expressions qualitative and less quantitative, making it difficult to integrate them into models of analysis.

Bibliography

1) [IVAN96a] Ivan, I., Sinioros, P., Popescu, M., Apreutesei, P., Dardala, M.: *Comparative Analysis of Software Reliability Models*, Proceedings of the Circuits Systems and Computers,

International Conference of Hellenic Navy, pp. 253-255, Athens, 15-17 July 1996
2) [IVAN96b] Ivan, I., Noşca, Gh., Pârlog, A.: *Asigurarea calității datelor*, Revista Asigurarea calității, nr. 8, 1996, pg. 8-15, Bucureşti
3) [IVAN06] Ivan, I., Noşca, Gh., Popa, M. - *Managementul calității aplicațiilor informatice*, Editura ASE, Bucureşti, 2006
4) [VOIN04] Voineagu V., Gradinaru G., Colibaba D., *Analiza statistică a datelor de mediu*, Editura ASE, Bucureşti, 2004
5) [VOIN07] Voineagu V., Țițan E., Ghiță S., Boboc C., Todose D. - *Statistică. Baze teoretice şi aplicații*, Editura Economică, Bucureşti, 2007
6) Assessing data quality with control matrices, Communications of the ACM, Volume 47, Issue 2, February 2004: http://portal.acm.org/citation.cfm?id=966389.966395&coll=portal&dl=ACM&CFID=7344227&CFTOKEN=88902836
7) Wireless sensor networks, Communications of the ACM, Volume 47, Issue 6. June 2004: http://portal.acm.org/toc.cfm?id=990680&type=issue&coll=portal&dl=ACM&idx=J79&part=magazine&WantType=Magazines&title=Communications%20of%20the%20ACM&CFID=7344227&CFTOKEN=88902836
8) Supporting data quality management in decision-making, Communications of the ACM, Volume 42, Issue 1, October 2006: http://portal.acm.org/citation.cfm?id=1217759&dl=ACM&coll=portal&CFID=7344227&CFTOKEN=88902836
9) The DaQuinCIS architecture:a platform for exchanging and improving data quality in cooperative information systems, Communications of the ACM, Volume 29, Issue 7, October 2004: http://portal.acm.org/citation.cfm?id=1024528&dl=ACM&coll=portal&CFID=7344227&CFTOKEN=88902836
10) Understanding user perspectives on biometric technology, Communications of the ACM, Volume 51, Issue 9, September 2008: http://portal.acm.org/citation.cfm?id=1378727.1389971&coll=portal&dl=ACM&CFID=7344227&CFTOKEN=88902836
11) Challenges and constraints to the diffusion of biometrics in information systems, Communications of the ACM, Volume 48, Issue 12, December 2005: http://portal.acm.org/citation.cfm?id=1101779.1101784&coll=portal&dl=ACM&CFID=7344227&CFTOKEN=88902836
12) Michael D. Kane, John A. Springer, Integrating bioinformatics, Distributed Data Management, and Distributed Computing for Applied Training in High Performance Computing, Proceedings of the 8th ACM SIGITE Conference on Information Technology Education, October 18-20, 2007, Destin, Florida, USA
13) Marcos Aurelio Domingues, Carlos Soares, Alipio Mario Jorge, A Web-Based System to Monitor the Quality of Meta-Data in Web Portals, Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, pp.188-191, December 18-22, 2006
14) http://www.invorad.com/InVoRad%20diode%20system2.pdf
15) http://linkinghub.elsevier.com/retrieve/pii/S0969805105002106
16) http://cat.inist.fr/?aModele=afficheN&cpsidt=17540332
17) http://www.freepatentsonline.com/6882949.html
18) http://www.studia.ubbcluj.ro/arhiva/abstract.php?editie=PHYSICA&nr=SPECIAL%20ISSUE&an=2001&id_art=2865
19) http://www.springerlink.com/content/w19104pn48g06822/
20) http://www.temarex.com/dataacq.htm
21) http://www.cs.cf.ac.uk/Dave/Vision_lecture/node12.html

22) http://www.techexpo.com/buyers-guide/SE-DA.html
23) http://www.freepatentsonline.com/7417427.html
24) http://www.microlink.co.uk/dataaq.html
25) http://zone.ni.com/devzone/cda/tut/p/id/7376