Information Integrity – A Structure for its Definition

by

Vijay V. Mandke
Director and Professor,
School of Engineering & Technology,
Indira Gandhi National Open University,
198, Asiad Village Complex, New Delhi-110049.

and

Madhavan K. Nayar
President,
Unitech Systems, Inc.,
1240 E. Diehl Road, Suite 300,
Naperville, Illinois 60563, USA.

Abstract

The paper addresses the research issue of integrity of information in computerized information systems. Specifically, the paper studies errors at different stages in the information system and argues that there are intrinsic Integrity attributes that all information systems must satisfy. These intrinsic Integrity attributes are identified to be Accuracy, Consistency and Reliability. The paper argues that errors due to factors drawn from system environment, external to the application system and overlapping the user environment, are not amenable to application controls conceived at system design stage, resulting in modification of data/information model to $<e, a, v + \eta>$ where η represents noise or error component that has occured but not corrected. This calls for automatic feedback control systems for on-line error detection (or estimation or predication as the case may be) for integrity improvement in information systems. The paper shows them to be sampled data control systems leading to Information Integrity Technologies and in the process presents a system's view of a structure for defining Information Integrity. The paper concludes by observing that continuing decentralization of high technology makes Information Integrity Technology an imperative for every information system, and the same will have to be developed in a computer language compatible with the information processing environment of the user organization. This opens opportunities for extensive research and product development and implementation initiatives in Information Integrity Technologies along the lines of its structure as developed in the paper.

1. Information System Model

While discussing data and information, Rajaraman observes that they are not the same [23]. Specifically, he points out, "data" is the raw material with which we start and "information" or "data product" is processed data which is used to trigger certain action or gain better understanding of what the data implies; thereby offering following relationship between "data" and "Information":

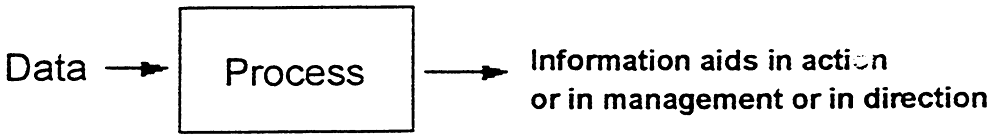
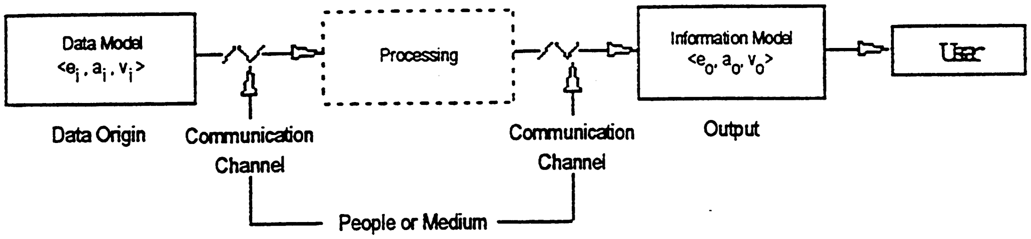


Figure (1) : Data-Information Relationship

What we are considering here is the Information System model that enables the contemporary organizations to act or to manage or to decide; thereby making processing of data to generate information for use a basic event. For studying integrity of such an Information System (IS) model, we need to further define "data" and various components or sub-systems of the IS model. In this context it is submitted that approach to defining data with an idealized view (i.e., a model) presented by a triple < entity, attribute, value > as developed by database research community offers a working method [9]. This is because this definition of data (which within the framework of IS model should also apply to information) treats data as a formal organized collection that can be studied. Further, it allows to segment integrity issues into issues concerning entities, attributes and values.

Further, with introduction of microcomputers and telecommunication into the computing process and with spread of data communication and particularly transaction processing networks with world wide reach, the communication channels at input and output ends are becoming important components of information system.

It is within the framework of above discussion, Figure (2) presents a conceptual model of an Information System.



Where $< e_i, a_i, v_i >$ denotes a triple for Data model and $< e_o , a_o , v_o >$ for Information Model

Figure (2) : Conceptual presentation of an Information System model.

2. Errors in the Information System

As observed earlier, information systems contain errors. These errors can be at various stages in the information system [13].

2.1 Data Origin Stage

This stage comprises designing and operating data collection procedures and codes, form filling (leading to data generation), data collection, and data preparation along with machine operation (where data is converted into machine-readable form). Various errors during operating data collection procedures and codes concern incorrect, incomplete and ambiguous manual, inappropriateness of manual language to user, non-availability of manual and carelessness, and they result into use of wrong procedure or code.

During filling in the data forms, errors encountered are in terms of ambiguous form filling directions, poor format, substitute or un-authorized person filling directions poor motivation and once again carelessness, resulting into incorrect filling of forms.

Coming to data collection stage, it is infected by errors caused by poorly designed forms and codes, poor handwriting and carelessness, resulting in omissions, inaccuracies, data in wrong place, loss of data, data manipulation, etc.

Data preparation stage along with machine operation stage where data is prepared and converted into machine-readable form suffer from errors caused by poorly written keypunch instruction, poor procedures, hardware errors, poor maintenance, misrouting of documents, sabotage, carelessness; resulting in incorrect operations, fraudulent operations due to crime, data not processed on time, and machine breakdown.

2.2 Communication Channel (Pre-Processing Stage)

Growth of on-line databases, distribution of information system resources across geographically separate locations, user community spanning entire enterprise, increased use of data communication functions, and information management networks combining terminal access to information processing and database resources, electronic mail, office automation, word-processing, fascimile, and graphics have understandably made "communication channel" a very significant component of an information system. And it so turns out that this communication channel is extensively prone to errors.

To begin with unlike long distance voice communications and conventional radio broadcast, data communication content and performance are affected by inherent signal interference, usually referred to as communication noise, thereby introducing errors in the data transmitted [19].

Other causes affecting telecommunication and thereby, introducing error in communication channel, relate to physical structure of telecommunication and to logical aspects of the data communication process itself. Specifically, causes relating physical structure of telecommunication (in addition to the factor of communication channel noise already mentioned) pertain to factors of : electromagnetic signal radiation, circuit switching deficiencies such as crosstalk between circuits and the failure of mechanical or electronic components of the switch itself, and any interruption in the electrical power supply used by the switching facility.

As regards to causes relating logical aspects of data communication, they include failure of the software used in either data communication network management or in data communication itself, and the size and inherent complexity of the network being used.

In addition to these, communication channel also suffers from problems of theft of service and circuit tapping, acts of sabotage, incidents of accidental destruction and of effects of adverse weather and water-caused damage.

2.3 Processing Stage

This operation transforming "data" into "information" comprises machine operation, use of data files, use of systems and application programmes, and of processing operation itself.

Errors during machine operation have been identified earlier. In respect of data files, errors caused are due to poor physical storage, lack of clearly defined responsibilities for data files, inadequate procedures, natural disaster and theft, fraud or sabotage; resulting in warped cards, dirty tape or disks, destruction of files, etc.

During the use of systems and application software, errors caused are due to out-of-sequence programming, wrong algorithms, wrong programming instructions, poor documentation, lax security; resulting in incorrect solutions and un-authorised changes.

Coming to the processing itself, perhaps the most significant cause for errors is carelessness in data processing; resulting in records lost, and use of incorrect file.

2.4 Communication Channel (Post-Processing Stage)

Errors and their causes in view of communication channel at post- processing stage are same as those for communication channel at pre-processing stage discussed earlier.

2.5 Output Stage

Finally, it is at the output stage that the user receives the "information" which is the output of the Information System. This "information" constituting "output" is used by the user as an aid in action or in management or in decision.

This "output" is the product of all input and processing under the Information System Model. Errors at this output stage are caused due to processing and operation errors; resulting in inaccurate and incomplete output.

2.6 Integrity Implications

Chamber's dictionary gives the meaning of integrity as "entireness, wholeness, unimpaired stage of anything, purity". Errors as above at various stages of an information system certainly result in loss of integrity at each of these stages as also in the loss of overall system integrity.

Specifically, errors at data origin stage, resulting into use of wrong procedures or codes, incorrect filling of forms, incorrect or fraudulent operations, data not processed on time, machine breakdown, etc., give rise to inaccurate, incomplete, backdated and insecured data, further threatened by loss of privacy in view of fraudulent operations.

Similarly, errors (during communication channel stage prior to processing) caused by communication channel noise, physical structure of telecommunication and failures in logical aspects of data communication, circuit tapping and theft of service, acts of sabotage, incidents of accidental destruction and the unpredictability of the complex networks used, give rise to inaccuray, incompleteness, loss of confidentiality and loss of privacy in data.

Further, during processing stage, errors in machine operation, errors in respect of data files, application and systems software and errors in processing itself, give rise to inaccurate, incomplete, insecured data further threatened by loss of privacy.

As pointed above, communication channel at post-processing stage also contributes to the loss accuracy, completeness, confidentiality and privacy - this time of processed data i.e., information.

Finally, errors at output stage also result in inaccurate and incomplete information. Before proceeding with the study of Information Integrity attributes, it may be mentioned that the literature deals with terms "Information Integrity" as also "Information Quality". Though integrity and quality account for implications of errors in information systems, in engineering production systems, the term "Quality" has a strong connotation of product uniformity, which is considered as the hallmark of the engineering production line. Against this data or information is not expected to be uniform in that each data or information is so because it is unique i.e., non-uniform. It is to accommodate such philosophical view points as also to emphasize wholism of data/information accessed and processed and to some extent maintain identify of studies in errors in information sytems from that of general engineering production systems, that the term Information Integrity may be considered more appropriate for study at hand in this paper.

3 Information Integrity Attributes : A Heuristic Treatment

As can be seen, errors in information systems result in loss of integrity at each stage of the information system, and, thereby, in the loss of overall system integrity. This integrity loss is in terms of the attributes (not to be confused with entity attributes referred to in this Section and paper) of : accuracy (purity), completeness (entireness, wholeness), data/information being up-to-date (i.e. timeliness implying accuracy inspite of time related changes in data/information), security and privacy (unimpaired meaning undamaged; purity).

Let us consider data/information modelled as a triple < e, a, v > as suggested in Section (1) above. As explained earlier, this affords a very meaningful approach whereby integrity attributes for an information system can be considered by studying them (integrity attributes) for the components of the triple, i.e., entity, attribute and value.

To elaborate, a universe for a company may comprise "employees", "products" and "customer orders". Employees and products represent "entity classes or types" and customer order represents "relationship" in the universe.

One can consider entity class, namely, employees class. As explained in Section (1), it may be represented by attributes as follows :

EMPLOYEE = (Employee Number, Name, Department Number, Salary, Date of Birth, Sex)

An entity is a specific occurrence of an entity type. Categories that make up an entity type are called attributes. The entity type shown above lists the following attributes : Employee Number, Name, Department Number, Salary, Date of Birth, Sex. Attributes either identify specific entities (for example, Employee Number), or specify particular facts or properties about entities (for example, Date of Birth).

Each entity attribute has a domain assigned to it; a domain is a set of permissible values. For example, the Salary attribute may have the domain from $ 10,000 through $ 20,000. In addition, one or more constraints can be imposed on admissible attribute values. For example, values for the Salary attribute may be subject to the constraint that the employee's salary can not exceed that of his or her superior.

Finally, a value provides information for specified attribute of a specified entity. For example, for employee entity, Albert, entity representation may be as follows :

Albert = (94256, Albert, 9, $ 15000 p.a., 6.5.75, M)

In such case, Date of Birth attribute has the value 6.5.75 and Salary attribute has the value $ 15,000 p.a. and so on.

With data/information model (in terms of triple) illustrated above, a clearer picture of integrity attributes could be obtained. Suppose the Information System Model (in Figure 2) delivers information about an organization - a company. Let "view" of this information comprise, as explained above, entity types "employees" and 'products" and relationship "customer orders". Further, let entity class, namely, employee class have view model as illustrated above.

Then to study accuracy of the information on the company, one may study accuracy of the information on entity types, namely, employees, products and on relationship customer orders. Further, to study accuracy of information on entity type employees, one may study accuracy of information about attributes corresponding to entity type employees. Finally, to study accuracy of information about attributes, one may study accuracy of values for attributes; thereby making the exercise of studying accuracy of information on the company a viable exercise.

In specific terms, given the employee entity representation, namely :

Albert = (94256, Albert, 9,15000, 6.5.75,M) the accuracy of information in respect of the entity type could be studied by studying accuracy of information about value of its each attribute. Above representation gives value of Date of Birth attribute 6.5.75. If the employee's Date of Birth is known, then it is possible to ascertain the accuracy of the information item from the information system. In this case, it is further possible to quantify the accuracy by computing the difference between the actual and the obtained Date of Birth.

However, suppose the information obtained gave values on the specific entity Albert as follows :

Albert = (94256, Albert, 9,15000, 6.5. --- ,M)

As can be seen, here in the value of the Date of Birth attribute, year is missing. Thus the information is incomplete. Incidentally, information, therefore, is also inaccurate. What thus emerges is for information to be accurate, it should also be complete; but every complete information is not necessarily accurate.

Let us consider another situation, this time concerning Salary attribute. Suppose Albert's Salary history is as follows :

Table (1) : Salary History for Entity Albert

Year	Salary $
1992	12000
1993	13000
1994	14000
1995	15000
1996	16000

If the information obtained for entity Albert as above is as of 1996, then value of $ 15000 pa.m for Albert's Salary attribute is not up-to-date and, hence, inaccurate for the year 1996, though the value was correct for the year 1995. This once again suggests that requirement of "timeliness" is also necessary for "accuracy", though not sufficient.

Finally, from the point of view of information model, requirement of security, meaning undamaged information, is analogous to accuracy, as any damage to information i.e., say to the value of an attribute will only result in inaccuray of the value.

Security also has an aspect of confidentiality. Further, security of information is also important from point of view of privacy. However requirements of confidentiality and privacy, though they emerge as implications of errors in the information system, can not be considered central requirements for all information systems, as there can be informations where confidentiality, and, for that purpose, security, and privacy may not be required.

Thus from the set of integrity attributes of accuracy, completencess, timeliness, security and privacy identified above, attribute of accuracy is central to an information system and attributes of completeness and timeliness are necessary for the attribute of accuracy. In other words, attributes of accuracy, completeness and timeliness are intrinsic to an information system irrespective of use of the information derived from the system. Against this, requirements of security in the sense of confidentiality and of privacy are optional to an information system and depend on the context and nature of use of information.

There are two other requirements that have not emerged in the integrity analysis so far and they are consistency and reliability of data/information.

Specifically, like completeness and timeliness requirements, consistency requirement is also a part of accuracy requirement, i.e., if data/information is accurate, then it is also consistent, but otherwise is not true. To clarify, consider the example of entity class employees with value for specific entity Albert as follows :

Albert = (94256, Albert, 9,16000, 6.5.75,M)

with salary attribute having domain through $ 10,000 to $ 20,000 and with Albert's Salary history from 1992 to 1996 starting with initial salary of $ 12,000 and ending with $ 16,000 with yearly increment of $ 1000 as given in Table (1) above. Further, there is, say, a constraint that Albert's Salary may not exceed his superior's salary which happens to be $ 17,000.

As can be seen, in the above example, value of Albert's Salary attribute, which is $ 16000, as also all values under his salary history are within the domain range of $ 10,000 to $ 20,000. Further, Albert's Salary attribute value

S 16,000 is less than his superior's salary value which is S 17,000. Hence information on Albert's Salary attribute value satisfies the domain as well as the constraint. Therefore, the information can be seen to be consistent.

Coming to the requirement of reliability its origin may be seen in the very choice of the Information System Model, wherein system output, i.e., information, is defined as what user receives as an aid in action or in management or in decision. Here no user is defined as such, but "utility" or "use" role of information is brought out. It is in this context, that the requirement emerges that information obtained be reliable.

To develop a perception about the reliability attribute, let us consider an example of an information system for processing of examination results for a distance learning activity. A view model for the activity may comprise entity types and relationship as follows : Courses, Course students, Study Centres and Course Examinations. For the relationship Course Examinations, the structure may be as follows :

Course Examination = (Course Number, Programme Code, Student Roll Numbers. Assignment Codes, Assignment Marks, Term-end Examination Code, Marks in Questions from Term-end Examination, Grade)

For our purpose, let us further consider the attribute "Marks in Questions from Term-end Examination", which consists of a domain comprising set of entities, namely, marks. Understandably, values of marks for each question has minimum and maximum which go to define the domain.

When a specific relationship, i.e., a specific examination for a specific course is considered under a given programme code, for each student (S_i), there would then be marks (m_{ij}) for each question (Q_j) for the attribute under consideration. This information item as obtained from the information system would then be available as follows :

Table (2) : Marks in Questions from a Term-end Examination for a Specific Course Examination.

Question Student	Q_1	Q_2	Q_j	Q_m	Total Marks
S_1	m_{11}	m_{12}	m_{1j}	m_{1m}	m_1
S_2	m_{21}	m_{22}	m_{2j}	m_{2m}	m_2
.	.	.	.	.	.
.	.	.	.	.	.
S_i	m_{i1}	m_{i2}	m_{ij}	m_{im}	m_i
.	.	.	.	.	.
S_n	m_{n1}	m_{n2}	m_{nj}	m_{nm}	m_n

The question at hand could be the reliability of values $\{m_{ij}\}$ for all i ε [i, n] and for all j ε [i, m] so that total marks obtained in Term-end Examination can be used along with Assignment marks to obtain for each student a grade giving reliable measure of student performances.

Various statistical methods are possible to study the above question; one such (method) at hand to explore integrity requirement of reliability being Analysis of Variance (AOV) technique where "R", reliability of information item obtained through values of attribute, namely, "Marks in Questions from Term-end Examination", is given by [12]:

$$R = \frac{\text{Variance of true values of marks } (V_T)}{\text{Variance of Information obtained on values of marks } (V_I)}$$

The value of R so obtained will always lie within the domain [0,1].

As mentioned earlier, marks for each question may have requirement that they lie within defined maximum and minimum under a domain. This requirement may be separately checked to ensure consistency of the information obtained from the information system on the marks for each question for a student. Of course, what is discussed above, is perceiving reliability as an accuracy with which the information obtained represents the data item in whatever respect the information system processed it and a methodology to quantify reliability for those attributes for which values lie on the real line. If it is not so and which often would be the situation in information systems one comes across, then one will have to once again look into the question of how reliability can be quantified (indeed so is the case with examples given earlier in respect of other integrity attributes, too).

In other words, immediate effort here is not to offer an all pervasive definition of reliability, but to state significance and centrality of such integrity requirement in respect of an information system. The attribute of consistency, being part of accuracy, also comes under this category.

Thus. taking entire discussion together. irrespective of the nature of use of the information obtained from the information system, attributes of accuracy, completeness, timeliness. consistency and reliability emerge as intrinsic attributes that an information system must meet, while attributes of security and privacy are optional depending on the context and nature of use.

There is more to the intrinsic integrity attributes mentioned above. As pointed out earlier, attributes of completeness and timeliness are necessary for accuracy. That is to say when checked for accuracy, the value of the information item also gets checked for its completeness and for it's being up-to-date (timeliness), as accurate value has to be complete and timely. In that sense, it is sufficient to check for accuracy only.

Similar is the situation in respect of consistency, too, as an accurate value also has to be consistent. However, difference is that, as illustrated earlier, consistency check is in terms of domain values and in terms of constraints without referring to real world objects and, therefore, a simpler and less expensive task offering first approximation on accuracy and, when checked in addition to accuracy, increasing overall reliability of integrity checking process itself.

It is within the above framework then Accuracy (includes completeness and timeliness), Consistency (satisfying constraints) and Reliability (accuracy with which information item represents data item in whatever way information system processed it) emerge as intrinsic or basic or objective attributes of Information Integrity. As mentioned earlier. what one is considering here is an Information System Model which delivers Information for use. Therefore. depending on IS application area and industry standards. without reference to any specific user there may be different requirements/standards of information. For example, although Accuracy of Information is intrinsic and central to the Information System performance, in many application areas irrespective of who is the individual user it is not uncommon for four digits to the right of the decional place to be rounded off without loss of much information. Thus, even though the intrinsic Integrity attributes of Accuracy, Consistency and Reliability emerge as basic Integrity attributes all information systems must demonstrate, depending on the application area, each of these intrinsic Integrity attributes may satisfy different application area specific industry standards.

As observed earlier, depending on the context and nature of use, there would also be user specific integrity requirements, namely, security and privacy and these emerge as extrinsic or subjective attributes of Information Integrity. There are other subjective attributes of Information Integrity, too, as reported in the literature based on user survey in respect of data/information requirements [18, 10]. These are : Usability, Independence, Precision, Relevance, Sufficiency, Understandability, Freedom from bias, Conciseness, Brief. Trustworthy, etc.

Just as in the case of intrinsic or objective attributes of Information Integrity, there is also a question of defining and, where possible, quantifying these extrinsic or subjective Integrity attributes. But then this query is beyond the scope of the query at hand in terms of deciding a structure for defining Information Integrity.

Figure (3) presents a system's view of the Information System Model in Figure (2), incorporating error implications and presenting emergent Integrity attributes.

4. Application Controls and Beyond

Traditionally, errors in information system (IS) and resulting integrity implications have been addressed as if they were preventable or correctable by building controls within application system. With time there have been efforts to put in greater inputs right at the system analysis and design stage, hoping that would ensure Information Integrity [13].

However, reality is different. Main reason for this is ironically the inadequacy of controls designed to meet lapses in integrity.

4.1 At different IS stages

4.1.1 Data collection procedures and codes

With shared data environments, with user community spanning entire enterprise and with human element of carelessness. it is extremely difficult to avoid situations like mannual unavailable when needed, use of unauthorized mannual. user inability to understand mannual language, resulting in the use of wrong procedure or code and. therefore, in inaccurate and insecured data inspite of controls.

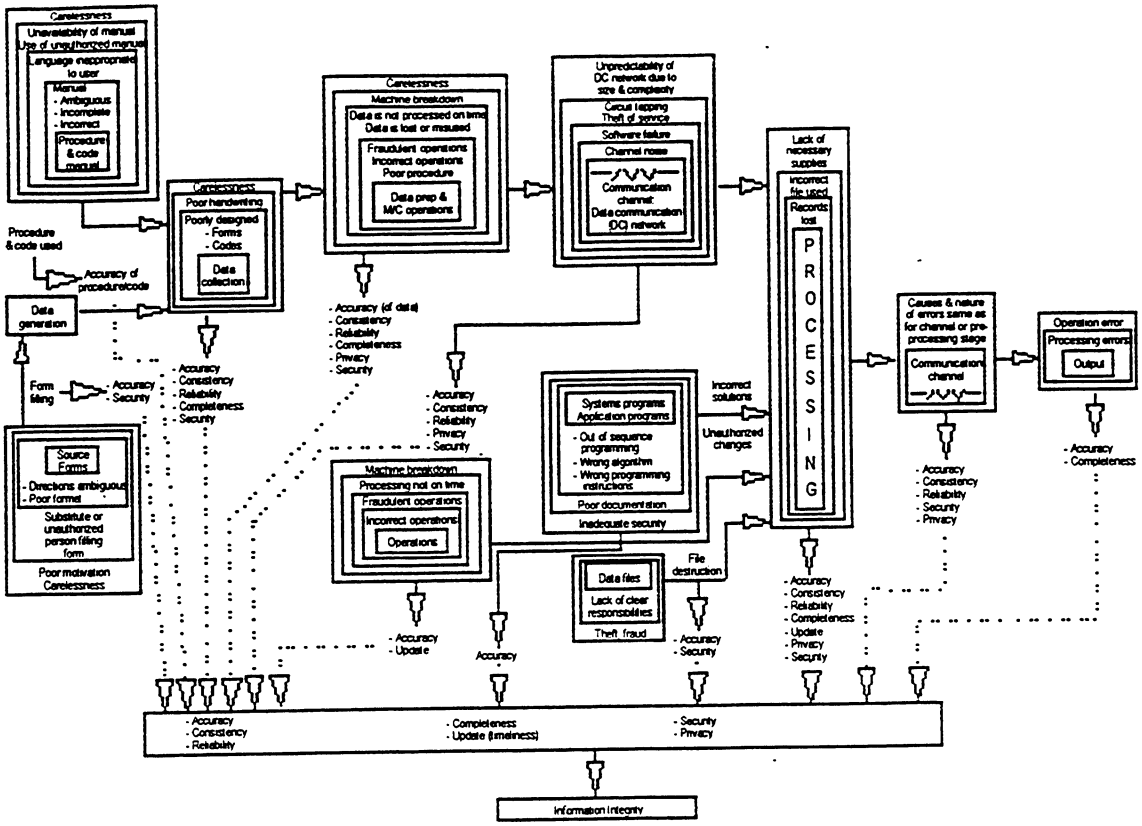


Figure (3) : System's view of the Information System model in Figure (2) incorporating error implications and presenting emergent Integrity attributes.

4.1.2 Form filling stage

Situations of poor motivation, carelessness and substitute or unauthorized persons filling out forms continue, resulting in incorrect filling of forms.

4.1.3 Data collection stage

At this stage, errors caused due to carelessness are very difficult to eliminate completely.

4.1.4 Data preparation stage

Controls at data preparation stage are not sufficient to eliminate error implications of carelessness in data preparation. Further, it is also extremely difficult to ensure controls like employee selection and training for foolproof results.

4.1.5 Machine operation stage

In respect of this stage, errors caused particularly due to human behaviour like carelessness of operators, sabotage, desire for personal gain, documents lost or misrouted are difficult to eliminate. Further, once again, it is difficult to have controls like upgraded personnel selection or training that are foolproof.

4.1.6 Communication channels at pre and post processing stages

Inspite of all communication controls, errors caused by failure of data communication software, channel noise, and unpredictability of data communication network due to its size and complexity, not to speak of adverse influence of people and weather, are difficult to eliminate.

Controls during machine operation at pre-processing stage and their inadequacy are same as that for machine operation following the data preparation stage discussed earlier.

4.1.7 Data files

Coming to data files. controls considered are controlled humidity storage, "clean room" conditions. special cabinets. periodic cleaning, centralized storage under librarian. upgrading of storage procedures. backup data and controlled access to files. However. these controls can not completely remove causes of theft, fraud or sabotage which stem from human behaviour.

4.1.8 Systems and Application Software

All controls in respect of systems and application software do not ensure removal of all errors. particularly those caused during programming and those due to poor documentation and lax security.

4.1.9 Processing stage

All controls at processing stage cannot remove errors that are caused by carelessness.

4.1.10 Output stage

Finally, controls at output stage can not always eliminate, all operation or processing errors.

4.2 Human error : a significant integrity problem

As one critically considers the above discussion, amongst others. factors of carelessness, poor motivation. and other actions of people emerge as most significant factors contributing to errors in information system and. thereby. lapses in integrity. According to a study performed by the Executive Information Network [6], 55% of the respondents involved in a survey considered human error as the most important integrity threat. Ironically, human error can also be one of the most serious problems causing system interruptions.

While the frequency of human error and the opportunity for human error are important considerations. the magnitude of the loss due to human error is also a major concern. Contrary to the general opinion of many information system practioners, human error is not always a low-consequence threat. In fact, in terms of money lost. human error is the largest single cause of economic and productivity loss in the information system integrity arena. As an evidence of this. consider a study reported in Computerworld [3], which attributes 52% of corporate information damage to human error. Single-incident losses can also be significant.

Rapidly evolving technologies producing distributed information networks and shared data environments that are common place today pose yet another issue. Once an error occurs it can be considered an "inherited error" [4] if it is passed along to another computer, network, database, file, or the like. Inherited errors occur when an error is propagated beyond the system in which it originated. For example. if a personal computer on a local area network is used to prepare a report - and erroneous data is incorporated into that report - when the report is submitted to another computer system, an error is inherited [36].

Various controls that are considered and implemented can not remove these human error possibilities. as can be seen from the feedback on error occurrences and their consequences as available from the field.

4.3 Inadequacy of controls

Information user's concern for integrity of information obtained from information system has undergone an almost cyclic transformation in the past few decades. Centrally located, batch-oriented systems of 30 years ago suffered from the inherent unreliability of their many discrete components, which made system failure frequent and maintenance complex and time- consuming. Consequently, much effort and ingenuity was expended in those days in devising fault detection and automatic correction techniques to defend computer against the effects of hardware failure to which it was prone.

Thus in the early days of computing, numerous studies were carried out on errors in data transcription. Experiments were devised where groups of people were assigned tasks of encoding, writing, copying, keying, and reading large volumes of data and results of these experiments were then analyzed to determine the frequency of various types of errors. Many commonly used data validation techniques and controls, including handwriting and forms design standards evolved on the basis of these findings [18].

Later, however, with the advances in manufacturing technology exemplified by medium scale and largescale integration. these problems were alleviated considerably, both by improved component reliability and an increase in density of logic gates per replaceable pack. which reduced the difficulty of maintenance. Concern over hardware unreliability and consequent information error therefore abated.

However, as computer systems become increasingly networked, as applications increasingly share data with one another, and as databases become increasingly distributed, the risk of data/information error - including risk of inherited error- increases, which amounts to the issue of data/information pollution.

4.4 Factors External to Application Controls

As discussed through this Section, these errors are caused by factors not amenable to controls including application controls conceived at system design stage itself. Of course, literature reports research efforts in terms of identifying foolproof information requirements [14, 20 & 33], but design experience shows this is something not easy to achieve.

This is because, due to factors detailed here, computerized information systems invariably have errors that are made but not corrected by the controls incorporated at system design stage. As can be seen, these factors, invariably have their presence mainly through the system environment which is external to computing (and hence the application) system and overlaps the user environment, though together they (the computing system and its external environment) constitute the Information System Model as in Figure (2). Inspite of application controls, it is these external factors that then make information systems give rise to information which is inaccurate, inconsistent and unreliable [21].

These external factors could be categorized into five major categories; namely, Change, Complexity, Communication, Conversion and Corruption.

Change

Change may occur either in the content or in configuration of the system environment, resulting in a possibility of error introduction in the information system. Every hardware change, software release, and organizational change will come under this category, offering cause for error and, therefore, for a possibility of an inaccurate, inconsistent or unreliable information.

Complexity

Whenever one introduces complexity, there is a possibility of error introduction in the information system. Every new component, be it a programme, database or network, adds new interfaces increasing the possibility of error introduction.

Communication

Communication stands for movement of data/information within or across enterprises and it also provides a chance for error introduction.

Conversion

Conversion, in this context, refers to the consolidation, decomposition or transformation of data. Whenever one converts data from one form to another, there exists a possibility of error introduction, resulting in information which may not be accurate.

Corruption

Finally, corruption pertains to human behaviour (poor motivation, desire for personal gain, carelessness, actions of people), to factors leading to inherited errors polluting the information systems, and to unpredictability (noise) of any kind leading to introduction of errors in computerized information systems.

Whether, in addition to controls discussed above, computerized information systems also incorporate human engineering design criteria at the system design stage itself or hardware and software vendors further incorporate error-checking filters into their products, it is these external error factors that then have to be addressed, if one were to resolve the question of errors in Information System Model so as to obtain Information which is Accurate, Consistent and Reliable.

5. Need for Automatic feedback Control System for on-line Error Detection and Integrity Improvement in Information Systems

In the information systems, the need is then to remove errors that are made but not corrected. When abstracted this implies, in the Information System Model in Figure (2), given that data/information is represented by a triple <e, a, v> and considering a particular example where say an output, i.e., processed data, i.e., Information is represented by Entity Class, namely, Employees and where specific entity (e) under consideration is an employee by name Albert and where specific attribute (a) under consideration is Albert's Salary, then, by virtue of on-line errors present in the information system, at any time, there exists a possibility of information item on value (v) of Albert's salary being inaccurate, inconsistent or unreliable, i.e. it's being affected by error or say corrupted by noise, and, therefore, a more realistic representation of value (v) is $(v + \eta \;)$, where η represents noise or error component.

It is within this framework of error implications on data/information model wherein triple $< e, a, v >$ is replaced by triple $< e, a, v + \eta >$ and, as discussed in Section (2), considering that these error implications are present at each stage of an information system; namely, data origin stage, communication channel prior to processing stage, processing stage, communication channel at post-processing stage and output stage, a modified version of a conceptual schematic of an Information System Model in Figure (2) emerges, accounting for errors that are made but not corrected. The same is given in Figure (4) below.

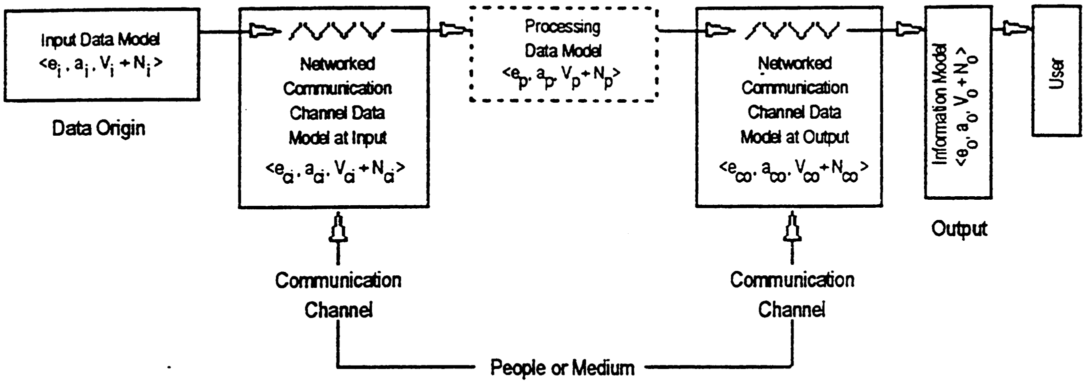


Figure (4) : Modified Conceptual presentation of an Information System Model in Figure (2) accounting for errors that are made but not corrected.

It is these on-line errors that are then to be removed. For this purpose, one has to first detect errors and then correct them. In other words, what is required is to incorporate on-line learning and error correcting mechanisms in the Information System Model. Specifically, this calls for automatic feedback control systems with error detection and correcting technologies for improved information Accuracy, Consistency and Reliability; technologies that maximize integrity of information systems - Information Integrity Technologies.

Rajaraman [24] points out that integrity of the overall information system as in Figure (2) is ensured if the integrity of all parts of the system are ensured (see Figure (5) below).

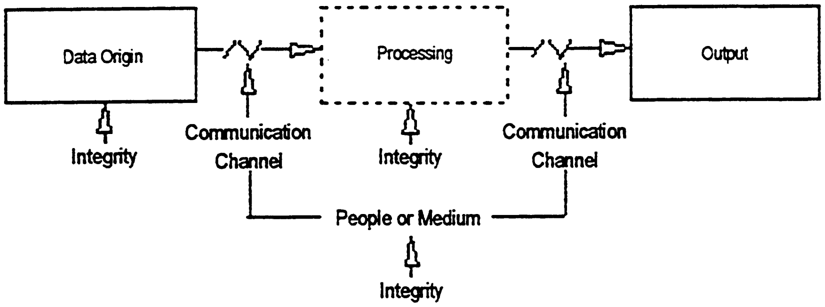


Figure (5) : Conceptual presentation of Integrity of an Information System

As a result, each stage and its intermediate stages in the information system as discussed in Section (2), along with the overall information system (with data as input and with processed data, i.e., information for use as output) considered as black box, become candidates for incorporating automatic feedback control systems or Information Integrity Technologies as above.

What Information Integrity Technologies will have to do is to follow data as it originates, moves over communication channel at pre-processing stage and gets processed, and follow processed data, i.e., information as it moves over communication channel at post-processing stage and gets used by user (as information system output), so as to detect where the error(s) occurs. Based on this, the Information Integrity Technologies would then need to take corrective action to remove error(s), so as to improve integrity of information system stage(s) under consideration and of the overall information system. In doing so, it would also be helpful if it is possible to measure integrity of a stage in the information system and of the overall information system, thereby offering a measure of integrity improvement achieved and a statement of improved level of integrity. Most importantly, such detection and correction of error(s) and improvement of integrity along with measurement of improvement and statement of integrity level achieved, will facilitate demonstration of integrity of computerized information systems rather than mere trusting them in the context.

However, apart from difficulties that are obvious in designing and developing such automatic feedback control systems leading to Information Integrity Technologies and in measuring integrity (i.e., Accuracy, Consistency and Reliability) of information, the most important difficulty in perceiving such technologies is that it is impractical to follow, i.e., track and analyse every bit of data/information for all times as it flows through the information system stages.

Way out here is to consider Information Integrity Technology that takes a sample of input data at the output or at an intermediate point of an appropriately identified stage or sub-system of the information system and then follows or keeps track of the sampled records at output or intermediate points of sub-sequent stages (sub-systems), at a given point of time or at different points of time over a required time intervel. Records so obtained at a given stage (sub-system) could then facilitate study of patterns of errors at that stage (sub-system) in the information system.

Utilizing the records available for different points of time over the time interval, error patterns can be studied even for a time variation. Based on patterns of errors for a given stage (sub- system) so analysed, causes of errors in the sampled data/information can be known so as to obtain corrective action to eliminate the causes, remove the error and improve the integrity. In other words, Information Integrity Technology could comprise sampled data control system for the stage (sub- system) under the information system as also for the overall information system.

As discussed in detail earlier, the on-line errors in information systems causing loss of integrity in information systems, are mainly due to factors drawn from system environment, external to the application system and overlapping the user environment. This in turn implies, it is users of information systems who have to do computer housekeeping by incorporating Information Integrity Technologies in the computerized information systems installed, so as to avoid on-line errors that are otherwise made but not controlled and corrected. Certainly this effort has costs, but then they have to be paid once it is decided to use computerized information systems, as Information Integrity Technologies introducing feedback, error correction and integrity optimization in information system performance are necessary (and sufficient) for Accurate (including complete and timely), Consistent and Reliable information which is the first responsibility of an information system.

6. Defining Information Integrity Structure

6.1 Reserach Survey [18, 22]

6.1.1 Integrity Model by Biba

One of the earliest attempts to build "Integrity" Structure (model) are due to Biba [5, 26]. Biba observed that multilevel security (MLS) policies are only concerned with controlling dissemination of information, while the validity of information in a computer system requires control over modification made to information. He called this later issues "integrity", and informally defined integrity in a computer system to be a guarantee that the system performs as intended by its creator.

His model further stressed the importance of correctness in notions of integrity. The Biba model of integrity excludes application issues, concentrating on properties of a system that would enhance assurances of integrity for a Trusted Computing Base (TCB) and for classes of data stored within the system.

6.1.2 IBM Concept of System Integrity

In 1981, IBM proposed a concept of "System Integrity" for its VM and MVS families of operating system [28, 29]. Specifically, IBM defined it as the provision of mechanisms under the customer's control which can prevent an unauthorized programme from : circumventing memory protection mechanisms, avoiding access control mechanisms and obtaining supervisor state.

Accordingly, issue of integrity of its operating system is the major theme of a set of guidelines that IBM makes to its customers [31] concerning aspects that must be considered when designing and implementing modifications or extensions to the operating system. This integrity support of IBM restricts only unauthorized (non-privileged) programmes, and IBM holds the customer responsible for access controls on libraries containing authorised programmes and the effects of any authorized programmes they develop.

O'Shea [22] observes that IBM offers a particularly narrow view of integrity. It is concerned primarily with the Trusted Computing Base (TCB) and in particular with the requirements that it be tamper-proof, that any access control mechanisms it may provide can not be circumvented and that it always be invoked.

6.1.3 Trusted Computer System Evaluation Criteria (TCSEC)

In 1983, the United States Department of Defence Computer Security Centre (now called the National Computer Security Centre) published the Trusted Computer System Evaluation Criteria (TCSEC). It was produced following an extensive programme of research into Computer Security requirements for military and government systems, with the intention of encouraging commercial developments of systems which would meet those requirements (at a reasonable cost). An evaluation scheme was established to certify products claiming to meet various levels of the criteria [11].

O'Shea points out that the concept of Trusted Computing Base (TCB) appears in Biba's integrity model and in IBM's "Statements of System Integrity", and it is fundamental at all levels of TCSEC. In other words, integrity issues in the TCSEC are thus concerned with correctness and protection of TCB [22].

6.1.4 The Network Interpretation (TNI) of Integrity

In 1985, the Network Interpretation (TNI) offered an interpretation of TCSEC that applies to networked systems [34]. In concrete terms TNI identified additional considerations that apply to networks, which were not explicitly addressed in the original TCSEC.

Thus TNI considers integrity issue in a system. The TCSEC aims at integrity primarily in terms of the need to protect mandatory security labels against unauthorised alteration, and what is referred to as "overall system integrity" (presumably the integrity of its TCB). The TNI introduces integrity requirements dealing with the need to ensure that information is reliably passed between components of the network. The issues introduced include correctness of message transmission, authentication of the source and destination of a message, and correctness of various data fields used to transfer user and protocol data.

The TNI observes that the term "integrity" when used in the context of computer systems has been used to refer to such issues as consistency, accuracy, concurrency, data recovery, modification access control and some concept of credibility or quality of information. The TNI concedes that integrity will often be important, equally if not more, as security (which we can interprete as "confidentiality" in this context), and considers integrity primarily as a property of a system that provides assurance as to the accuracy, faithfulness, non- corruptibility and credibility of information transmitted between source and destination entities.

In this interpretation, an integrity policy addresses both intentional attempts to modify information (referred to as Message Stream Modification), and the unintentional but largely inevitable threat to transmitted information that occurs through noise in communication systems and equipment failure. An integrity policy thus places great emphasis on the ability to write objects, and constrains modification of information to comply with the integrity policy [22].

The TNI recognizes that the mechanisms used in support of integrity policy requirements are often probabilistic in nature, and states that the desired probability with which message stream modification can be detected should be specified in the integrity policy.

TNI requires that support for the integrity policy in a network system, which largely depends on communication protocols, must be provided by a Network Trusted Computing Base, which must ensure that the mechanisms supporting the integrity policy are protected and always invoked.

6.1.5 Clark-Wilson Integrity (CWI) Model

In 1987, Clark and Wilson presented the Clark-Wilson model of Integrity [7, 1]. Since then, this model has provoked constructive reaction from within computer science fraternity and academia. A number of papers have been published that discuss or propose solutions to its requirements, and the United States Institute of Science & Technology (NIST, formerly the National Bureau of Standards) has held invitational workshops to encourage that development and discussion of its basic ideas [22].

CW Integrity model argues that the TCSCE does not adequately address the security issues predominant in commercial environments. It argues that while military environment is more concerned with issues of information confidentiality, commercial environments are primarily concerned with preventing un- authorised modification of data as a countermeasure to risks of error and fraud. CWI Model also suggests that countermeasures to these threats have long been established as part of common business practice, principally in the concepts of well-formed transactions and separation of duties, neither of which is directly supported by the TCSEC.

Clark and Wilson, thus suggested that separate policies are required for confidentiality and data integrity, and that, with the exception of common requirements such as user authentication, much of the mechanism for supporting these two - security and integrity - policies would also be different [22]. In the process, the Clark-Wilson Integrity (CWI) model can be described in terms of its framework for integrity requirements comprising following five stages [1]:

- Stage 1 : Trust objectives
- Stage 2 : External-Interface requirements model
- Stage 3 : Internal Requirements Model
- Stage 4 : Rules of Operation
- Stage 5 : Functional Design

The CWI model captures a fundamental view of integrity for trusted system, with emphasis on data integrity. As a result of the deliberations that have taken place since their original paper was published, Clark and Wilson have refined their view of the concepts and mechanisms involved, but still stand by the principles of their original model [35]. Their view of data integrity has been refined to recognize that it involves internal consistency of a system (a correctness aspect) and external consistency (correspondence with reality), in turn requiring development or improvement [22].

6.1.6 Terry and Wiseman view of Integrity

Terry and Wiseman have in 1989 presented a "new" model of security in a paper [32] that provides some illuminating observations on the nature of the Clark-Wilson notion of the data integrity in particular. They argue that there are two different notions involved in the Clark-Wilson "model" of data integrity.

The first notion, for which they reserve the term "integrity", is concerned with the internal correctness of a system, and can be used to prevent users from causing effects for which they are not authorised. This is the "error control" objective of the Clark-Wilson model, and is supported by their TP mechanism (which Terry and Wiseman suggest is an issue of Typing).

The second notion is concerned with correspondence to the real world, which they call "appropriateness". This is concerned with whether an action is desirable or "right" thing to do, irrespective of whether it is authorised, and addresses the Clark-Wilson objective of fraud control. They to recognize that correspondence with the real world is difficult to model or to enforce, but accept that separation of duties is an effective mechanism in practice.

6.2 An Analysis

6.2.1 Lack of precise Information Integrity attribute definition

With military requirements dominating the research in information systems, the issue of secured computer systems and of confidentiality of information has always been a high priority query. As a result, for over twenty-five years, there have been efforts to work on information security programmes. Further, security has normally been taken to mean confidentiality, integrity and availability [27, 8]. Most people involved with information security issue are at ease with this terminology except that the meaning of word "integrity" is not adequately resolved, the word being frequently used, as mentioned earlier, to describe a range of attributes (or requirements) such as : quality, correctness, usability, relevance, prevention of unauthorized modification, protection, appropriateness, etc [22].

6.2.2 A critical look

However, a critical look at the research investigations reported above bring forth the following important aspects :

Security and Integrity are different

a) There is a recurring realization that security policies are only concerned with controlling dissemination of information (confidentiality), while there is also the problem of the validity of information in a computer system-requiring control over modifications made to information (also termed as integrity or correctness of information) [5, 7].

Separate Mechanisms needed for Security and Integrity

b) In fact CWI model suggests that separate policies are required for confidentiality and data integrity, and that with the exception of common requirements such as user authentication, much of the mechanism for supporting these two - security and (data) integrity - policies would also be different [22].

Notions of Correctness and Appropriateness

c) There is also an appreciation that there are two different notions involved in CWI model of "data integrity": (i) one of "integrity" concerned with internal correctness of a system and (ii) other of "appropriateness" concerned with correspondence with the real world [32].

Assumption of Trusted Computing Base and Procedures

d) Further, concept of Trusted Computing Base (TCB) appears to be fundamental to Security and Integrity. Specifically, CWI model postulates that security and integrity are dependent upon [16]:

- trusted computer base;
- trusted procedures;
- trusted processing;
- authenticated procedures and audit trails; and
- segregation of duties.

Network Integrity

e) In the context of today's technologies, telecommunication networks are integral to a computer-based information system. From this angle, research investigations reported concede that integrity will often be equally important, if not more, as security (taken to mean confidentiality). Accordingly integrity is primarily considered as a property of a system that provides assurance as to the accuracy, faithfulness, non-corruptibility and credibility of information transmitted between source and destination entities (it may be mentioned that this view of network integrity again relies on the concept of Network Trusted Computing Base).

Noise in the Information System

f) Finally, in the context of (e) above, while considering integrity issues in respect of networks, it is further realized that the unintentional but largely inevitable threat to transmitted information occurs through noise in communication systems and equipment failure. As a result, the mechanisms used in support of integrity policy requirements may need to be probability with which message stream modification can be detected being specified in the integrity policy.

6.3 A reiteration of Heuristic understanding of Integrity

When the observations in Sub-section (6.2) above, are seen against the heuristic view of integrity implications of errors in Information System as developed in Section (3), it clearly emerges that intrinsic or basic or objective Information Integrity attributes of "Accuracy" and "Reliability" go to make or correspond well with attributes of "correctness" and "appropriateness" of information as recurringly identified in the Information Integrity research reported [see observations 6.2.2 (a), (c) and (e)]

Further, it also emerges that, consistent with the position taken in Section (3), the research investigations reported also view the issue of integrity as separate from that of security [see observations 6.2.2 (a) and (b)].

6.3.1 Towards a precise and agreeable definition of Information Integrity

It is within this framework then the Section (3) proposition that argues "Accuracy" and "Reliability" as intrinsic Information Integrity attributes gets further reiterated. Indeed, the heuristic study of Section (3) also defines the attribute of "Consistency" which is a necessary condition for "Accuracy", simpler and less expensive to check; in turn offering a viable first approximation on Accuracy and, when checked in addition to Accuracy, increasing overall reliability of integrity checking process itself.

Thus Section (3) proposition presenting Accuracy, Consistency and Reliability as intrinsic Information Integrity attributes that an Information System must meet gets further confirmed, in turn offering a precise and agreeable definition of Information Integrity.

6.3.2 Inadequacy of assumption of Trusted Computing Base and Procedures

The observations on the reported research investigations also mention the assumption of Trusted Computing Base and control of input for integrity [see observation 6.2.2 (d)] and recognize the presence of "noise" in an information system [see observation 6.2.2 (f)]. In regards to the requirement of the Trusted Computing Base and Control over inputs. it may be mentioned that Report of IFIP Working Group 11.5 [16] points out that this is a narrow view of what constitutes integrity and it is confined within the logical bounds of an information system as it excludes the material impact of the people and business application processing necessarily involved in any system. In fact. the report goes to state that "the concept that a system can be trusted over time without the ability to provide the evidence that the trust is well placed is incompatible with internal control principles. The concept of trust therefore is insufficient for our purposes".

A similar observation is also made by Courney and Ware in their paper 'What do we mean by Integrity' when they point out that "a 1991 study conducted for the National Research Council explains integrity as assuring that information and programmes are changed only in a specified and authorized manner, which regrettably ignores the need for integrity in hardware, the physical environment, and people" [8].

6.3.3 Information System error implications due to Noise and factors external to the application system

And in regards to observation 6.2.2 (f), wherein there is a recognition of "noise" in information system. it only exemplifies. in addition to justifying the concept of intrinsic Information Integrity attributes identified irrespective of the nature of use of Information, the issue of information system error implications due the factors external to the application system.

6.4 Nature of Information Integrity Technology

Thus both these observations. namely, that of inadequacy of assumption of Trusted Computing Base and of recognition of noise in an information system only go to substantiate the analysis of Section (4) on inadequacy of application controls due to external factors of change, complexity, communication, conversion and corruption, and of Section (5) on recognition of presence of on- line errors that are made (noise) but not detected. and, thereby, of need for on-line automatic feedback control technology, i.e., Information Integrity Technology, for continuous error detection and integrity improvement.

6.4.1 Information Integrity Attribute Quantifiers

Having defined Information Integrity attributes and identified the nature of Information Integrity Technology, in its effort to develop Information Integrity Structure, the investigation at hand may consider, at this stage the issue of quantifying the level of integrity. In turn, this calls for quantification of degree of Accuracy, Consistency and Reliability.

6.4.1.1 Accuracy

Accuracy refers to correctness, i.e., preventing unauthorized modification, i.e., degree of conformance between a particular value of data/information and an identified source. The identified source provides the correct value [9]. It can be an object or relationship in the real world; it can also be the same value in another database. or the result of a computational algorithm.

Given that value of data/information is expressed in a numerical, Accuracy of the data/information can be quantified in a number of ways [25, 9, 30, 2]:

i) Difference between the actual value (i.e., value of the identified source) and the value processed by the information system.

ii) Error Ratio $= \dfrac{\text{Actual Error}}{\text{Acceptable Error}}$

iii) Accuracy Index $= \dfrac{\text{Number of correct values}}{\text{Number of total valaues}}$

iv) Number of records examined : R

Number of records with atleast
one defect of loss of Accuracy : D1

$$\text{Percent Defective} = \left[\frac{D1}{R} \times 100\right]$$

$$\text{Accuracy Index (A)} = \left[1 - \left(\frac{D1}{R} \right) \right]$$

Note : Percent Defective is a quantifier used extensively in statistical quality control.

v) Number of defects (cases of loss of accuracy) detected : D

Number of records examined : R

$$\text{Defects/Losses of accuracy per record} = \frac{D}{R}$$

$$\text{Accuracy Index (A)} = \left[1 - \left(\frac{D}{R} \right) \right]$$

It may be mentioned that defect denotes accuracy violation, i.e., presence of error, and hence the absence of accuracy. Ratios based on defects/errors can be converted into accuracy ratio by the transformation:

Accuracy Ratio = 1-Defect (i.e., Error) Ratio.

Understandably notion of Accuracy quantified as above has many issues not considered here. What if correct value of the identified source is undefined, or simply unknown. And of course what if data/information is say a name or has an alpha-humeric value or is a video image; how is error or defect defined then ?

6.4.1.2 Consistency

Consistency is with respect to a set of constraints. As pointed out earlier, data/information is said to be consistent with respect to a set of constraints if it satisfies all constraints of the data/information model [9]. Constraints can apply to the same attributes in different entities (such as the salary attribute in the entities of several employees); they can also apply to different attributes in the same entity (such as the salary level and salary attributes in the entity for a particular employee).

Given the number of constraints specified (CS) and given the number of constraints for which error/defect detected in the sense constraints are not satisfied (CE), then consistency can be quantified as follows [30]:

$$\text{Consistency (C)} = \left[1 - \left(\frac{CE}{CS} \right) \right]$$

6.4.1.3 Reliability

Finally, as mentioned in Section (4), Reliability (R) may be considered as an accuracy with which the information obtained represents the data item in whatever respect the information system processed it. For this purpose, a model may be considered where any processing of data has a large error component, random in nature. As a result volume of error in the processed data will be different each time the data processing is repeated, leading to significantly different information in each case; thus reflecting a low reliability of the information. Thus "Reliability" refers to the extent of existence of random errors in an information, or in other words, the degree of consistency with which an information can be repeated, without any intervening or additional instruction.

Coming to the quantification of reliability (R), in any data/information model, for an entity (i), the value (v_i) for an attribute or processed value for ith data item for the entity may be expressed as $v_i = t_i + e_i$, where 't_i' is the true component of the value and 'e_i' is the error component. It is assumed that :

(a) v_i takes values on a real line.

(b) e_i's are distributed independently and randomly over the whole population of data items (i's) and that $e_i = 0$, and

(c) e_i's are uncorrelated with t_i's.

Then reliability 'R' is given by :

$$R = 1 - \frac{V_e}{V_v}$$

Where

$$V_v = \frac{1}{N} \sum_{i=i}^{N} (v_i - \bar{v}_i)^2$$

is the variance of the processed value and

$$V_e = \frac{1}{N} \sum_{i=i}^{N} (e_i - \bar{e}_i)^2$$

is the variance of the error component.

From above it follows that reliability "R", also termed as "Coefficient of Reliability" or "Reliability Index", will have a value between 0-1.

It is appreciated that it may not be possible to repeat every data processing. In such case internal consistency of a data/information set comprising (a) information from processed data, (b) information from relevant identified source, (c) information from another related data base, (d) results from relevant computational algorithm, etc. could be studied to obtain the reliability.

Various methods exist for calculating the Reliability Index (R); Analysis of Variance (AOV) technique mentioned in Section (3) being one such. Choice of a method would depend on advantages, disadvantages and convenience of application in a given situation, while accounting for factors like nature of available data, form of data and computation aids available for processing.

With integrity attributes quantified as above, a measure for Information Integrity which encompasses these (and therefore more than one) attributes can be arrived at by taking recourse to analytical techniques applicable in studying multi-variable or multiple-objective situations particularly encountered in Education while assessing professional personality traites of a learner or in examination results under a course-wise passing system. Thus with integrity attributes quantified as above from an user end, an Information Integrity Profile based on attribute index values can be developed.

For the the purpose of further quantification, depending on the range between [0–1] in which an attribute index value falls, the attribute can be assigned a 5-point scale H – D with numerical points [5 – 1] attached to scales in that order. Further, each application area, consistent with information usage requirement, can have application area specific order of significance for integrity attributes. Let W_A, W_C and W_R represent significance weightages for the integrity attributes Accuracy, Consistency and Reliability, respectively, for the application area under consideration. These weightages may take values between [0-10].

In such case based on scales obtained for the attributes' index values between [1–5] for the given user application at hand and based on corresponding attribute signficance values between [0–10] for the application area for which the information system is designed, a specific Cummulative Information Integrity Index (CIII) can be defined, thereby offering on overall and holisitic integrity measure for the information system under study. It is this value of CIII and the values of the scales for the attributes as available from the Integrity profile that would enable the user to state his Integrity improvemnet need, inturn providing a basis to design and develop Information Integrity Technology for the purpose.

Before one proceeds with further development of Information Integrity Technology structure, a word of caution is warranted here. The quantification of integrity attributes is not a trivial task even when it is possible [25] and quantifiers suggested above do not bring out the complexity involved. In respect of Accuracy quantification, it is already mentioned that there could be a problem of correct value of the identified source (also called standard) being undefined, or being simply unknown. In situation an assumed standard itself may be incorrect as is often the case with data gathered some time in the past and with no corroborating evidence. In yet another situation there may be more than one correct value. Then there is a problem of how to quantify accuracy if the value does not lie on a real line, i.e., it is not a numerical. As regards to Consistency quantifiers, though a relatively simpler concept than Accuracy, it can assume complexities when all real data base inconsistencies are to be measured (and which will be the need) or when Consistency is also to be studied for the conceptual view of the data or information. Finally, as already mentioned, Reliabililty quantifier gives an index of an accuracy with which the information obtained presents data item in whatever way the information system processed it. There can be no one way of calculating the Reliability index and there will always be a need to develop one based on nature of available data, form of data and computation aids available for processing. All these areas then constitute the further research needs in the context of Integrity attribute quantifiers for Integrity improvement.

6.4.2 Information Integrity Technology Structure

With a suggestion for Cumulative Information Integrity Index (CIII) as above, within the framework of Information Integrity attributes of Accuracy (A), Consistency (C) and Reliability (R) argued, one can then consider defining Information Integrity Structure for on-line error detection and control or more specifically for integrity improvement role.

Figure (4) gives a conceptual representation of an Information System Model accounting for noise, i.e.. errors that are made but not corrected by controls built in at system analysis and design stage of the Information System. As observed by Svanks in her significant work entitled "Integrity Analysis : A methodology for EDP Audit Data Quality Assurance" published in 1984 [30], an Information System could be viewed as a production line in a manufacturing environment. Processing stage represents logic steps which utilize input transactions as raw material or parts to yield processed data base records, i.e. information, as the end product. A typical production line incorporates process control, but more importantly, also employs product control. The identification of faulty processes alerts product quality control to invoke special procedures such as tightened inspection, repair or discarding of finished goods. Conversely, the disclosure of substandard products suggests remedial action for specific processes.

Information System testing becomes the equivalent of process quality control in that the errors revealed call for software revisions and maintenance. If the Information System is already operative, a test result indicating error would only suggest that such error may have occurred in the past. Test procedures do not access the production data base, therefore, no statement can be made as to whether or not this Information System error has occurred in real environment. nor which records have been affected by the erroneous process. Therefore, as argued in Section (5) on "automatic feedback control system for on-line error detection and integrity improvement in Information System" the confirmation of potential or suspected anomalies on a live data base and subsequent integrity improvement becomes an essential facility (beyond application controls) within an Information System.

And the users of computerized information systems have to undertake this computer housekeeping to incorporate this facility in their information systems, so as to avoid potential serious losses occasioned by errors that were made (due to factors external to application control) but not corrected. In concrete terms this facility, constituting the Information Integrity Technology, will be an application and user specific software which, for an Information System Model as in Figure (6), on-line periodically and systematically samples records arriving at an appropriately chosen point (in the Information System Model) and then follows or keeps track of sampled records at subsequently identified points through the information system and stores the records so sampled and obtained through follow up (audit trail), to set up error detection database which is then analyzed to identify errors, i.e., changes not expected (irregular changes) and to quantify resulting loss of integrity therefore, followed by integrity improvement action wherein Information Integrity opportunity is identified and implemented.

6.4.2.1. Information Integrity Development Steps

R&D in Information Integrity Technology would thus follow steps given below :

i) Understand the user application of the computerized information system under consideration.

ii) Based on application area and based on organizational practices studied.establish organizational standard pertaining to data/information with reference to requirements of : Accuracy, Consistency, Reliability and Cumulative Integrity.

iii) Study data/information flow through the Information System and define database(s).

Note : Apart from knowing how the Information System processes the data and apart from understanding more about the "noise" in the system, the study would also necessiate knowing wherefrom, how and data/information of what integrity flows into the system.

iv) Based on the understanding of data/information flow in the system, for the database identified, develop the Information System Model as in Figure (6).

v) Specify and document the data rules, also known as edits, to be implemented to study accuracy and consistency of the data/information.

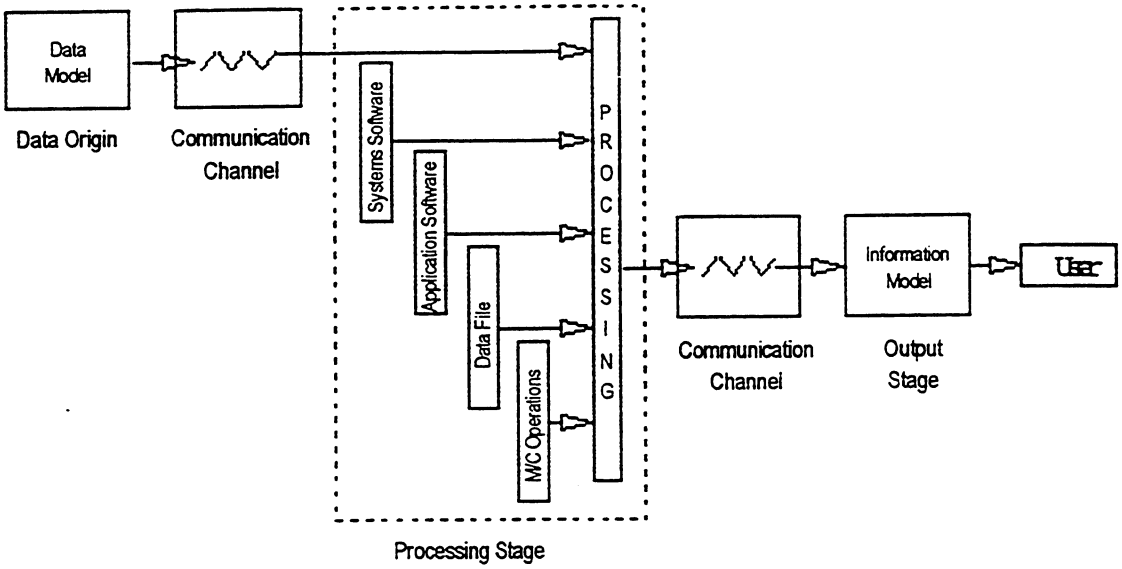


Figure (6) : Data/Information flow model for an Information System for developing Information Integrity Technology

vi) While accounting for factors such as nature of available data, form of data and computation aids available and keeping in view advantages, disadvantages and convenience of application, choose a method for calculating Reliability Index.

vii) Develop Integrity Analysis Software for analysing intrinsic Information Integrity attributes of Accuracy, Consistency and Reliability.

In addition Integrity Analysis Software may also undertake statistical analysis through time series analysis (and such other techniques) of error patterns (signifying irregular changes) contributing to loss of Accuracy and Consistency and of causes contributing to loss of Reliability; in turn leading to developing :

a) a filter to detect error or cause that occured sometime in the past at time $(t - \tau)$,

b) an estimastor to estimate error or cause that occured in the immediate past at time (t), and

c) a predictor to predict error or cause that may occur sometime in future at time $(t + \tau)$.

viii) For the Information System Model in Figure (3), select a data sampling point at the output of a subsystem (or at an intermediate point within the subsystem), as close to the beginning of the Information System Model as possible.

ix) Depending on how data arrives at the sampling point (continuously or in batches), develop a continuous or batch processing sampler (a sampling programme) to randomly select a sample or records arriving at the sampling point. Along with sampling records, the sampler programme should also select some identifier of the sampling point and record of the data and time of sampling.

x) Following the selection of a sampling point and development of a sampler, select points for maintaining audit trail for sampled records.

xi) These points for maintaining audit trail may be selected at points at the output of subsystems (or at intermediate points within the sub-systems) following the sampling point.

xii) Once the points for maintaining audit trail for records sampled are identified, develop a Sampled Records' Audit Trail (SRAT) programme to separate or pull out (at the points selected) the audit records comprising details such as : identifier for the record collection point, identifier for the person(s) entering the transaction records, transaction record number, date and time of record collection, inter-day sequence for transaction,

transaction type. transaction details. file change "image" consisting of both the before and after records. etc.

xiii) Ensure that sampler programme and SRAT programme so developed can download sampled records and records for audit trail as in (ix) and (xii) above into a database to be set up (see Step (xiv) below).

xiv) Accordingly. based on hardware and software considerations and based on number of sampled and audit trailed records. download the sampled and audit trailed records on mainframe or minicomputer or on personal computer/workstation so as to set up an Error Detection Database.

xv) Using the Integrity Analysis Software developed in (vii), analyse the Error Detection Database to :

a) identify data rule violations in respect of Accuracy and Consistency attributes,

b) based on data rule violation statistics, establish degree of integrity of data/information in respect of Information Integrity attributes of Accuracy and Consistency,

c) obtain Reliability index for the database along with analysis of factors contributing to the level of Reliability,

d) based on indices for Accuracy. Consistency and Reliability attributes. develop Integrity Profile and Cumulative Information Integrity Index, and

e) study changes in database not expected. i.e.. irregular changes.

xvi) Compare the Integrity profile and indices obtained as in [(xv(b)) - (xv(d))] with standards in (ii) - local. regional. national. international as the case may be - and with the user specifications on Integrity. so as to know what is expected of Information Integrity Technology. This would also facilitate ordering or ranking the Integrity attributes from the point of view of which attribute needs maximum improvement effort.

xvii) Then. for each of the Integrity attributes of Accuracy and Consistency, by further analysing the irregular changes either by subsystem or by field (in that order of priority of choice) locate separate Integrity improvement opportunities at each of appropriately identified pairs of a given field at a given subsystem.

xviii)Similarly based on Reliability Factor analysis in (vii), locate Reliability improvement opportunities at each of subsystems.

xix) Having located pairs of a given field at a given subsystem each for improvements of Accuracy and Consistency and having located given subsystems for Reliability improvement opportunities. further analyse the Error Detection Database and study irregular changes at each of pairs corresponding to each of Accuracy and Consistency attributes and study Reliability Factors at each of the subsystems. so as to understand over the time error patterns and causes contributing to loss of Accuracy, Consistency and Reliability.

This would then facilitate detecting error or cause that occured sometime in the past $(t - \tau)$. or estimating error or cause at time (t), or predict error or cause that may occur at a future time $(t + \tau)$.

xx) Based on assessment as in (xvi) of Integrity improvement target and based on the understanding of error patterns and factors for loss of intrinsic Information Integrity attributes as in (xix), now develop Information Integrity Improvement Action Plan for locations identified in respect of Integrity Improvement opportunities. This Integrity Improvement Action Plan may comprise restructuring subsystem(s) previous to the point of occurrence of error. improving integrity of data origin stage. improving communication channels, etc.

xxi) Finally. study performance of the Information System on incorporation of the Information Integrity Technology as outlined above. Accordingly obtain the intrinsic Information Integrity attribute indices. Integrity profile and Cumulative Information Integrity Index and compare them with appropriate reports before implementation of Information Integrity Technology available vide [xv(b)], [xv(c)] and [xv(d)], so as to quantify Integrity improvement achieved and to check if it is as per customer expectation.

6.4.2.2 Standards for Product Development

The Information Integrity Technology is needed for every Information System and will have to be developed in a computer language compatible with the information processing environment of the user organization. In other words. the associated documentation will have to be made available to many parties - IS designers. planners. operational personnel and management. Therefore, it must be unambiguous, concise and readable for the organization implementing the Information Integrity Technology.

This calls for organizational IS planning, devising policies, standards, and guidelines pertaining to data. If this is not ensured, net result is non-compatible, and hence unsharable data/information. The diverse IS within an organization may employ different element lengths or values for the same types of data/information. For instance, name and address conventions may vary, and account status may be specified as 1 (active), 2 (inactive) and 3 (deceased) in one IS and as A, I and D in another. This lack of standards introduces high overhead in many areas including that of development of Information Integrity Technology.

As can be seen, data rule specification is a very important step in development of Information Integrity Technology. In specifying data rule once again the first step is defining data rule standard. Understandably, the data rule standard is also required for undertaking Information Integrity Analysis.

Yet another area that calls for standards pertains to degree of integrity. As mentioned earlier, the application area would influence the requirement of how much Accuracy or Consistency or Reliability. The application area would also influence values of W_A, W_C and W_R. Further, quantification of Integrity attribute such as Accuracy calls for identification of data/information sources and their standards, i.e. correct values. In development of Information Integrity Technologies, it would therefore be necessary to establish these application area specific standards representing requirements of degrees of integrity as also of values of Integrity attribute significance factors.

Standards are also called for documentation. As can be appreciated, clear, concise and standard documentation of data rules constitutes the single most important source of Information for the programming task. The documentation, if recorded in machine-readable form, also serves as input to the reporting facility.

Finally, it is important to appreciate that development of standards as above would facilitate development of Information Integrity Technologies for different subsystems of the information system as also for the total system, and call for support of reputable software developers and vendors for the purpose; thereby inturn opening a new vista in terms of design, development, commissioning, operation and maintainance of data technologies, hitherto not attended, for ensuring on-line integrity of computerized information systems. These Information Integrity Technologies would cover data/information in various forms - numerical or alphabetic or alpha-numeric or video-images or any other - and that for different application areas and, as mentioned, depending upon the Integrity Improvement opportunities, would be applicable to data origion stage, networked communication stage, machine operation stage, system software and applilcation software stages, data file stage, processing and output stage as also to the entire information system, and would comprise such of the components and capabilities as mentioned below :

- Standardized data rule list,
- Standard sampling software,
- Standard Sampled Records' Audit Trail (SRAT) software,
- Standard sample file and error file,
- Standard statistical runs,
- Standard reporting facility,
- Standard filters, estimators and predictors,
- Standard support data base to preserve integrity analysis results,
- Standard mechanisms for deciding and implementing Integrity Improvement Action Plans, etc.

It is these requirements of standards alongwith the Information Integrity Technology Development Steps presented above, then provide an aggreable, precise and working proposition for defining the Information Integrity Structure. This structure of Information Integrity revolves around the intrinsic Integrity attributes of Accuracy, Consistency and Reliability. As mentioned in Section (3), it is appricated that there may be requirements of extrinsic or subjective attributes of Integrity as seen from the user point of view. Figure (7) presents a conceptual schematic giving a system's view of the Information Integrity Structure developed incorporating both - intrinsic as well as extrinsic - attributes of Integrity.

7. Conclusion :

Errors in computerized information systems were relatively manageable as long as there was homogeneous system environment and centralized control over information. Emerging trends of globalization, changing organizational patterns, strategic partnering, electronic commerce and distributed computing have changed all this, posing risks to Accuracy, Consistency and Reliability of information. These intrinsic Information Integrity attributes are central to

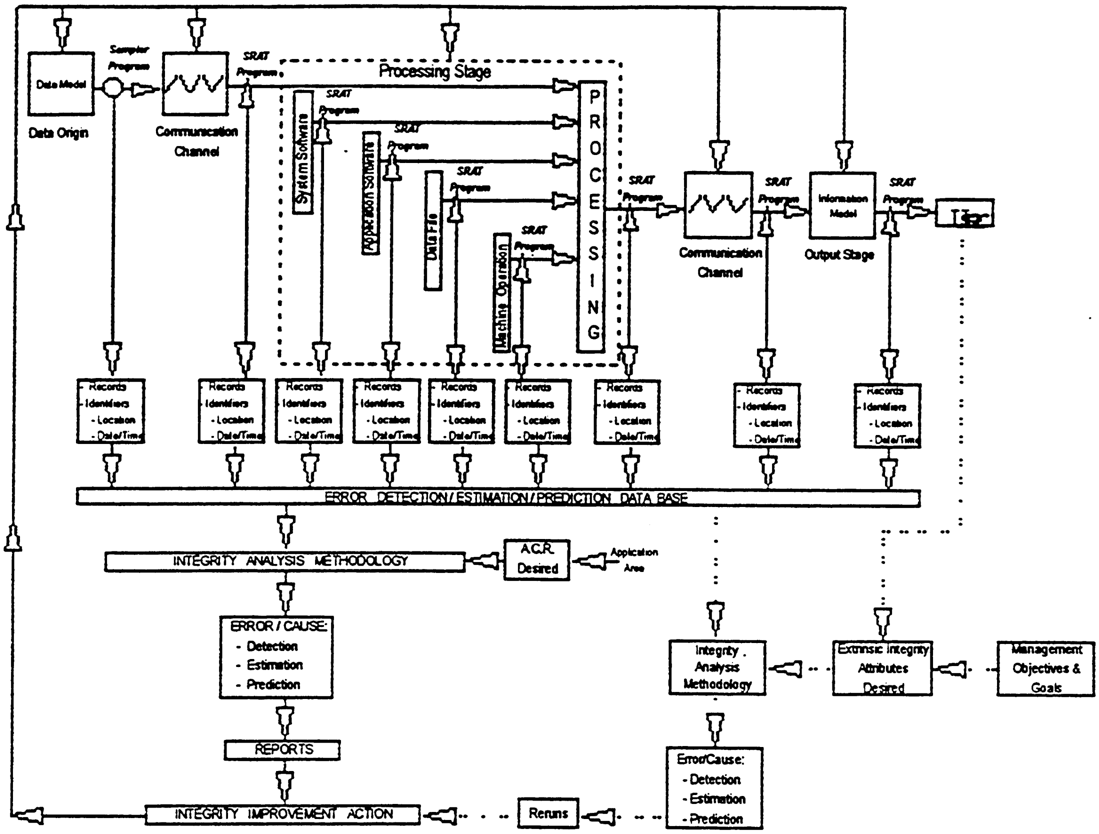


Figure (7) : Conceptual schematic giving system's view of a structure for defining Information Integrity and its technology.

any information system in that in their absence the information systems will have massive amounts of polluted (error-filled) data and useless, even dangerous information.

These errors are essentially caused by on-line factors of Change, Complexity, Communication, Conversion and Corruption which have their presence mainly through system environment which is external to computing (and hence the application) system and overlaps the user environment. Inspite of application controls, it is these external factors that then introduce, in information systems, errors that are made but not corrected.

Need therefore is to design and develop automatic feedback control systems leading to Information Integrity Technologies that (a) carryout on-line detection (filtering problem), estimation and prediction of errors contributing to loss of Accuracy and Consistency and of causes contributing to loss of Reliability, and (b) implement Integrity Improvement Action Plan accordingly. It is such Information Integrity Technologies that could then also provide a measure (metric) of Information Integrity achieved, inturn facilitating demonstration of integrity of computerized information systems rather than merely trusting them in the context.

It is appreciated that there are difficulties in developing such automatic feedback control systems leading to Information Integrity Technologies, the most important difficulty being how to track and analyze every bit of data/information for all times as it flows through the information system stages of : Data Origin. Communication Channel (pre-processing stage), Processing, Communication Channel (post-processing stage) and Output stage. Way out here is to develop Information Integrity Technologies as Sampled Data Control Systems.

Literature extensively refers to requirement of security in information systems. However, security and integrity are different and need separate mechanisms. Further, the concept of 'Trusted Computing Base and Procedures' implying an information system can be trusted over time without ability to provide the evidence that the trust is well placed is incompatible with internal control principles and therefore insufficient for the question of error free computerized information system. Further with networks becoming integral to information systems, there is need to recognize presence of "noise" in data/information models; thereby conceiving probabilistic descriptions of information flow models and of on-line Integrity Improvement mechanisms.

It is within above framework of Information Integrity Technology that Integrity attribute quantifiers and measures for Integrity profile and Cumulative Information Integrity Index can be developed followed by Information Integrity Technology development. The resulting Information Integrity Technology is then a SOFTWARE PRODUCT consisting : user data rule list for error detection, Integrity Analysis Software, the sampling programme, the Sampled Records Audit Trail (SRAT) programme, the programme for statistical analysis and for detecting, estimating and predicting errors/causes, the generation of Error Detection Data Base and its analysis based reporting, deciding and implementing Information Integrity Action Plan (probabilistic and manual action plans included) and documentation.

By the very nature of errors in information systems, the Information Integrity Technology is needed for every information system and will have to be developed in computer language compatible with the information processing environment of the user organization. In other words, associated documentation will have to be unambiguous, concise and readable for the user. This calls for development of standards, thereby calling for support of reputable software developers and vendors for the purpose.

Finally, Information Integrity Technologies as above would cover data/information in various forms and for different application areas and, depending upon Integrity Improvement opportunities, would be applicable to different subsystems of information systems, thereby inturn opening (along the lines of Information Integrity structure presented here) a new vista in terms of development of body of knowledge in this area of Data Science and in terms of design, development, commissioning, operation and maintenance of data technologies, hitherto not attended, for ensuring on-line integrity of computerized information systems.

References

1. Abrams M.D., Amoroso, E.G., Lapadula, L.J Lunt T.F., and Williams J.G., "Report of an Integrity Research Study Group", Computers & Security, 12 (1993) pp. 679-689.
2. Ameen D.A., "Systems Performance Evaluation", J. of Systems Management, (March 1989), pp. 33-36.
3. "Annon. Data view : loss leaders", Computerworld, (May 1987).
4. Banks W., and Weimer J., "Human Factors in Computer Systems", Prentice Hall Publishers, NY, (1992).
5. Biba K.J., "Integrity Considerations for Secure Computer Systems", USAF Electronic Systems Division, Bedford, MA, (1977), ESO-TR-76- 372.
6. Bloombecker B., "Commitment to Security", National Centre for Computer Crime Data, Santa Cruz, California, (1989).
7. Clark D.D., and Wilson D.R., "A Comparision of Commerical and Military Computer Security Policies", Proc., (1987), IEEE Symp. on Security and Privacy, IEEE, New Yark, (1987), pp. 184-194.
8. Conrtney R.H., and Ware W.H., "What Do We Mean by Integrity", Computers & Security 13, (1994), pp. 206-208.
9. "Data Quality Foundations", Published by AT&T Quality Steering Committee, U.S.A., (1992).
10. Delone W.H., and McLean E.R., "Information Systems Success : The Quest for the Dependent Variable", Information Systems Research 3(1), (1992).
11. "Department of Defence Trusted Computer System Evaluation Criteria", Department of Deference Computer Security Centre, Fort George G. Mead, MD, (1983), CSC-STD-001-83.
12. Guilford J.P., "Psychometric Methods", Tata McGrow Hill, India, (1978).
13. Hussain Donna, and Hussain K.M., "Information Resource Management", Richard D. Irwin, INC., Homewood, Illinois, (1984).
14. Kliem R.L., "Back to Basics : Developing A Good Requirements Document", Jornal of Systems Management, (October 1992), pp. 16-19.
15. Langefors B., "Theoretical Analysis of Information Systems", 4th ed., Philadelphia : Anerbach Publishers, (1973).

16. List W.. and Melvilse R. "Integrity in Information Systems - Executive Summary : IFIP Working Group 11.5". Computers & Security, 13. (1994), pp. 295-301.

17. Lochovsky F.H., and Tsichritzis. "Data Models". Engiewood Clitts. NJ: Prentice Hall. (1992).

18. Mandke Vijay V., "Research in Information Integrity : A survey and Analysis", Proceedings of the JNCASR and SERC Discussion Meeting at IISc Campus, Bangalore on Information Integrity - Issues and Approaches, Edited by Rajaraman V. and Mandke Vijay V., published by Information Integrity Foundation, New Delhi, India, (1996).

19. Menkns Belden, "Understanding Data Communication Security Vuluerabilities" Computer & Security. 9, (1990), pp. 209-213.

20. Mostert D.N.J., and Solms S.H. Von, "A Methodology to include Computer Security, Safety and Resilience Requirements as part of the user requirements", Computers & Security, 13 (1994), pp. 349- 364.

21. Nayar M.K., "A Framework for Achieving Information Integrity", Proceedings of the JNCASR and SERC Discussion Meeting at IISc Campus, Bangalore on Information Integrity - Issues and Approaches. Edited by Rajaraman V. and Mandke Vijay V., published by Information Integrity Foundation, New Delhi. India. (1996).

22. O'Shea G.F.G., "Operating System Integrity". Computer & Security, 10 (1991), pp. 443-465.

23. Rajaraman V., "Analysis and Design of Information Systems", PHI, New Delhi. India, (1991).

24. Rajaraman V., "Information Integrity - An Overview", Proceedings of the JNCASR and SERC Disucssion Meeting at IISc Campus, Bangalore on Information Integrity – Issues and Approaches. Edited by Rajaraman V., and Mandke Vijay V., published by Information Integrity Foundation, New Delhi. India. (1996).

25. Redman T.C., "Data Quality : Management and Technology", Bantam Books. NY. (1992).

26. Schroeder M.D., Clark D.D. and Saltzer J.H., "The Multics Kernel Design Project", ACM Operating Systems Rev., 11(5), (1977), pp. 43- 56.

27. "Security Functionality Mannual", DTI Commerical Computer Security Centre, V.21-Version 3.0 (unpublished).

28. "Statement of MVS System Integrity", International Business Machines, (October 1981), IMB ULET ZP 81-0801.

29. "Statement of System Integrity for VM/System Product Release 3 and Companion VM/SP High Performance Option Release", International Business Machines, IBM Programming Information Customer Letter, (June 1983 IBM ULET 5664-147, 5664-173).

30. Svanks Maija I., "Integrity Analysis : A Methodology for EDP Audit and Data Quality Assurance". EDP Auditors Foundation, Inc., (1984).

31. "System Programming Library" : System Macros and Facilities, Vol. 1, International Business Machines. pp. 193-204.

32. Terry P., and Wiseman S., A "New" Security Policy Model", Proc. 1989 IEEE Symp. on Security and Privacy, IEEE, NY, (1989) pp. 215- 228.

33. Tompkins F.G., and Rice R. "Integrating Security Activities into the Software Development Life Cycle and the Software Quality Assurance Process", Computers & Security, 5 (1996), pp. 218-242.

34. "Trusted Network Interpretation of The TCSEC", The National Computer Security Centre, Fort George G. Meade, MD, (1985), DoD 5200 28-STD.

35. Wilson D.R., "The Integrity Initiative - Where are the Auditors ?".COMPACS '90, Proc. 14th Int. Conf. on Computer Audit, Control and Security. The Institute of Internal Auditors. U.K., (March 1990).

36. Wood C.C., and Banks W.W. Jr, "Human Error : An Overlooked but Significant Information Security Problem", Computer & Security, 12 (1993), pp. 51-60.