

INFORMATION INTEGRITY KNOWLEDGE DEVELOPMENT INITIATIVE – AN INTRODUCTION

(A CIIR white paper)

Vijay V. Mandke
Research Leader,
Center for Information Integrity Research,
Unitech Systems India (Pvt.) Ltd.
B-64 (First Floor), Gulmohar Park, New Delhi-110 049, INDIA
E-mail: ymandke@unitechsys.com

PREAMBLE

Database research community has always acknowledged impact of data errors on Databases, and accordingly there are concepts of Data Security and Data Integrity. However, Date distinguishes between the two by observing that while security involves ensuring that users are *allowed* to do the things they want to do, integrity involves ensuring that the things they are trying to do are *correct*. In other words, security means protecting the database against *unauthorized* users, whereas integrity means protecting it against *authorized* users (direct as well as indirect). Indeed, integrity, unlike security, is applicable even in a single-user system (it is always desirable to avoid errors); and it is far more relevant when the system is shared.

1. LACK OF PRECISE INFORMATION INTEGRITY DEFINITION

Literature reports different definitions of integrity structure. Traditionally, researchers working in the area of Electronic Data Processing (EDP) have been looking at integrity issues. Specifically, in the early days of computing, numerous studies were carried out on errors in data transcription. Many commonly used data validation techniques, along with handwriting and forms design standards, evolved on the basis of these findings. Another popular research area has been the formulation of check digit algorithms, supported by verification studies with real numbers. These studies yielded confidence levels for the various algorithms.

Initial application of this research direction is to be seen in accounting literature that places emphasis on internal systems and audits. Indeed, as early as 1968, Feltham identified accuracy, timeliness and relevance as the three dimensions of data quality and analyzed their relationship with the value-in-excess-of-cost criterion within the context of the data consumer. The current literature on data/information quality considerably expands this list of integrity attributes to include requirements of: accuracy, usability, reliability, independence, timeliness, precision, completeness, relevance, sufficiency, understandability, freedom from bias, consistency, trustworthy, brief, etc.

With military requirements dominating the research in information systems, the issue of secured computer systems and of confidentiality of information gained a high priority query. As a result, for over forty years, there have been efforts to work on information security programs. Further, security has normally been taken to mean confidentiality, integrity and availability. Researchers involved with information security issue are at ease with this terminology *except* that the meaning of the word "integrity" is not adequately resolved, the word being frequently used to

describe a range of attributes (or requirements) such as: validity of information in a computer system (Integrity Model by Biba, 1970); prevention by customer of unauthorized program from: circumventing memory protection mechanisms, avoiding access control mechanisms and obtaining supervisor state (IBM concept of System Integrity, 1981); preventing unauthorized modification of data (Lipner's Integrity Application, 1982); correctness and protection of Trusted Computing Base and Procedures (Trusted Computer System Evaluation Criteria (TCSEC, 1983); reliability, accuracy, faithfulness, non-corruptibility and credibility of information transmitted, protection from noise in communication systems and equipment failure, consistency, accuracy, concurrency, data recovery, modification access, credibility of information (The Network Interpretation (TNI) of Integrity, 1985); internal consistency of a system (a correctness aspect) and external consistency (correspondence with reality – an appropriateness aspect) (Clark-Wilson Integrity (CWI) Model, 1987); property of data which have not been subject to unauthorized alteration or destruction (Integrity Model of International Standards Organization); property of internal correctness of a system - error control objective, and property of correspondence to the real world – fraud control objective (Terry and Wiseman view of Integrity, 1989); correctness and consistency of data and their classification labels (Sea View Model of Integrity); etc.

2. AN ANALYSIS

2.1 Integrity and Security are different

A critical look at the research investigations reported above shows that there is lack of precise Information Integrity attribute definition. Also, there is a recurring realization that security and integrity are different in that the 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 (termed as integrity or correctness of information). In fact CWI model, which has provoked constructive reaction from within computer science fraternity and academia, 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. There is also appreciation that integrity has two different notions involved: that of correctness and of appropriateness.

2.2 Inadequacy of assumption of Trusted Computing Base and Procedures

Further, concept of Trusted Computing Base (TCB) appears to be fundamental to above security and integrity definitions. However, literature in information systems research argues 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, IFIP deliberations go 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 (so mandatory to safeguard assets from misappropriation and ensure reliable information systems)".

2.3 Shift from Information Technology to Information

And this visualization of integrity issue acquires a further research dimension, as, today,

the reality of data driven technology (keyed to the flow of information across the enterprise and on the Net) and real time pressures of achieving business objectives of effectiveness and efficiency (through integration maximization) make continuous automation of ‘informational work’ (carried out by the soft components of the enterprise) a means for strategic and competitive advantage in a complex and changing environment. Specifically, this research issue dwells on modeling the generic business process (and each of its activities) *emphasizing* ‘information’.

This introduces a very significant dimension in the research query pursued in that the challenge shifts from one of dealing with information technology to that of information. Information is at different levels. At one level, information refers to the potential message in an entity or event, or in reports about it. Information is viewed, in this context, as ‘data’, i. e., as a function of the source only. At another level, it refers to transmission of message; as a function of both source and means of conveyance as in communication theory, which uses probability to quantify the properties of symbols to convey message. Its primary value in studying management information systems lies in the key ideas of probability and reduction of uncertainty and notions of noise, lag and error in transmission. And at a higher level, information refers to the meaning gained by the recipient (human as well as machine); the extent to which uncertainty is reduced and recipient knowledge increased. Information, in this context, is a function of source, process (includes communication medium and/or people) and the specific recipient.

When systems were seeking to produce only ‘standard’ product in high volumes with emphasis on operational and cost efficiency for strategic advantage, information requirements were seen to be structured and periodic and mainly at first two levels. However, (a) with the availability of on-line computers; computers providing capability of moment by moment optimization of processes and decision-making; and system integration capability (so as to yield a computer integrated system for system objectives), and (b) with technological reality of digital data as medium of information flow across the enterprise so as to ‘put it all together’, it has now become possible to use information at the third and higher level ‘smarter’ so as to achieve strategic and competitive advantages of larger mass-customized markets and financial optimization.

What this calls for is modeling a system – and a system can be described by a generic business process covering the entire supply chain from concept to delivery and follow up – as an *IS* processing unstructured and non-periodic information — a structural variant from the days of systems seeking ‘standard’ products. At the core level, this essentially facilitates viewing every activity (whatever else it does) as an *IS* activity (a cybernetic view), wherein ‘data’ is seen as raw material, ‘data transformation or processing’ as a system function characterized by pre- and post-processing communication channels (which include medium and/or people), and ‘data product’ or ‘information’ as processed data used to trigger information *use*, so as to deliver ‘information decision’ for controlling the physical process characterizing the activity. In other words, ‘data’ and ‘information for *use*’ are different entities, and database is seen together with data acquisition and information utilization cycles.

3. IS MODEL – AN INDIVIDUAL INFORMATION ORIGINATING AND PROCESSING SITUATION

The research investigations on “Information Envelope and its Information Integrity Implications” by Mandke and Nayar presented at International Conference on Information Quality 2001 [1] studied the problem of modeling a generic business process as an integral to a closed loop information and control system. Most information processing involves some type of data conversion to information in *use* and, therefore, is closely related to a decision process with an objective. Even when the information is transmitted without changing form, as in a communication system, the issue is to decide the purpose or objective of the transmission.

Traditionally, within the system-engineering framework, decision process is viewed to comprise of stages of forecasting (prediction), evaluation of alternatives and selection; information being considered basically as function of “source” (i.e. data) and at the most of “source” and “process”. However, an information and control system which is characterized by an unstructured and a periodic information processing is an open system; information being function of “source”, “process” and “recipient” (i.e. customer). For it more workable model of a decision process spans multiple stages. They are: based on long term goal set, *obtaining* ‘many factors’ & ‘multiple criterion’ characterizing problem (task) complexity; from multiple criterion, *recognizing* (deciding) on operable goal; from operable goal statement, *defining* planning & design constraints and opportunity spaces; from ‘many factor’ information variables characterizing problem complexity, *culling out* useful (relevant) information variables; *recognizing* relationships (interdependencies) between culled out information variables; *developing* state transition models defining dynamic behavior of culled out state (information) variables; and undertaking customized planning & design for *generating* alternatives for evaluation and final selection of flexible information decision for control implementation.

For details of development of this multiple stage decision process *IS* one may refer to the paper by Mandke and Nayar mentioned above [1]. What is significant here is to observe that *all* of the above stages, from obtaining ‘many factors’ and ‘multiple criterion’ and from operable goal setting to final choice of flexible information decision for control implementation involve *information originating and processing* activities with reference to their respective information bases - a structural variant from traditional view of decision process, which is concerned only with alternatives and information that are *already* “generated” and does not anticipate (or in the manner of a closed system, rather, has no need for) “origination” of information and “generating” alternatives. In other words, traditionally, decision process has been defined in the gamut of alternatives identified “*exogenous*” to the *situation* of decision-making and, hence, is a “collective” decision process. Against this, in the wake of open system based *IS* view of a system, the decision process has a requirement to identify operable goal, originate information and generate alternatives, all, “*endogenous*” to the *situation*, thereby making it (decision process) a candidate for designating it as an “individual” decision process.

Thus, what we have before us is a generic business process *IS* comprising of multiple stage “individual” decision process - that is an *IS* as an *individual information originating and processing situation*. Certainly, as with traditional information system which is a “collective” decision process, the “individual” information originating and processing situation is also characterized by physical plant/process/activity level uncertainty due to input noise, process parametric noise, and measurement noise, thereby introducing information errors and, hence, loss of Information Integrity. But, of still greater implication is the reality that at each decision stage

these information originating and processing activities are further affected by uncertainties due to the system environmental factors of 5“C”s, namely, complexity, change, communication, conversion and corruption; resulting in errors in information processed from stage to stage.

4. IS MODEL – A CONTINUOUS INDIVIDUAL INFORMATION ORIGINATING AND PROCESSING SITUATION IN THE PRESENCE OF UNCERTAINTY

More specifically, system environmental factors act on an *IS* externally as well as internally. Multiple decision process stages mentioned above, along with the control implementation stage (production, process and physical variable controls included) and the physical plant/process/activity stage, go to make an all-encompassing generic business process *IS* view covering manufacturing, production and service sectors. As mentioned above, at the physical plant/process/activity stage and at the physical variable control stage, there are system environment based operational uncertainties due to input noise, process parametric noise, and measurement noise. Then at all the control levels, which in the wake of “application” emphasis are so dominated, there are uncertainties due to information overload, lack of standardization, lack of relationship in data in several applications, and due to errors in hardware, software, data entry, or accidental or intentional failures, etc. These uncertainties result in information errors leading to loss of Information Integrity.

Further at managerial decision level controls there are uncertainties due to incomplete knowledge of system dynamics and due to judgmental errors at human-*IS* interfaces. In fact higher-level controls like say production control, apart from human-machine interfaces, in all probability, may even include humans as part of the process to be controlled. All this only adds to uncertainty, resulting in information errors and, therefore, leading to loss of Information Integrity.

And then we have the multiple individual decision process stages mentioned above, which it so works out are impacted by the system environmental factors of 5“C”s. Briefly stated, the entirety of the business process *IS* view described above makes it a complex *IS*. To begin with, at the operational, i.e., at the physical plant/process/activity level itself, this complexity introduces hitherto unknown complex error mechanisms coming from system development and implementation life cycle phases and that, too, coming with delay. Further at the stages of operational level and at control levels, there are errors that arise due to failures of embedded systems. Then, in the presence of emphasis on system integration maximization, there is uncertainty at all business *IS* view stages due to the presence of system interfaces, which expose innermost system modules to uncertainties due to external system environmental factors and vice versa. All this introduces only *further* errors and loss of Information Integrity in business process *IS* view.

And, finally, particularly the decision process stages (from obtaining ‘many factors’ & ‘multiple criterion’ and from operable goal setting stages to customized planning & design for evaluation of alternatives and selection of flexible information decision stage) involve information processing under multiple goals (implicit goals included), many factors, and a large number of interdependent information variables, varying with time, and not completely and correctly observable, and its system dynamics is not well understood; resulting in information errors and loss of Information Integrity.

Once again for a detailed statement of uncertainty in business process *IS* view under consideration here, one may refer elsewhere [1]. For the purpose of present investigation, however, two points need attention. Firstly, business *IS* view as here involves obtaining ‘many factors’ & ‘multiple criterion’ and operable goal setting rendering it “individual” decision process. This reinforces treating information processing under business *IS* view as individual situation involving information origination. Secondly, there comes the question what if the “goal” leading to usefulness factor with reference to information originated, though given, continuously needs adjustment due to constantly changing environment (effect of 5“C”s) (as very well can be the situation in say a service sector) or is not known or is out of date or is by itself complex (all these are the conditions to be observed in the real world problem solving). Even if one takes a conservative view, a large, semantically complex, time-pressured, tightly coupled, high consequence, high-reliability engineering system, in the wake of unclear goal statement (implicit goals inclusive), is observed to run a risk, in the fashion of an open system, of taking a life of it’s own.

In such case then the tasks of culling out the relevant facts (usefulness factor with reference to data and information variables) and of defining their interrelationships under the subsequent decision stages of the business process *IS* view cannot be treated as static ones determined uniquely and exogenously as in case of closed systems, but would acquire dynamic - open and endogenous in that – characters in the presence of 5“C”s, and they (data and information variables) would need to be *continuously* originated and processed.

This modeling reality has an implication for the *IS* modeling exercise underway as what it does is to model information processing under the generic business process *IS* view as a *continuous individual information originating and processing situation in the presence of uncertainty*, so as to account for demands of continuously determined specific goal based individual situation in a complex and changing environment. Further, as argued, this business *IS* view at all its stages is affected by information errors and by resulting loss of Information Integrity; thereby further increasing the requirement of ensuring the dependability or trustworthiness of information, i.e., Information Integrity.

5. TOWARDS A WORKABLE DEFINITION OF INFORMATION INTEGRITY ATTRIBUTES

Within the framework of above issues - of cybernetic view of an activity, and of requirement that information is for *use* - then a workable definition of intrinsic Information Integrity attributes can be considered and they can be defined as: Accuracy – the degree of agreement between a particular item (value, algorithm, configuration) and an identified source (standard); Consistency – satisfying constraints, i. e., the degree to which multiple instances of the item satisfy a defined domain; and Reliability. These are intrinsic or objective or basic Information Integrity attributes in the sense all information systems must demonstrate them. As information is for *use*, depending on *IS* application area and independent of any specific user, there will be different requirements/standards of information. It is understood that the intrinsic attributes would satisfy these different application area specific industry standards. Further, it is suggested that for data/information model represented by a triple $\langle e, a, v \rangle$, this identification of intrinsic attributes

is workable in that it allows segmenting integrity issues pertaining to those concerning entities, attributes and values.

Coming to the security, it may be seen to have two aspects; namely, undamaged information and confidentiality. Security implying ‘undamaged’ is synonymous with the requirements of accuracy or correctness, and to that extent is synonymous with the intrinsic requirement of accuracy, which is seen as integral to the definition of Information Integrity. As regards to the confidentiality aspect, it can also be seen from the point of view of ‘privacy’. It is submitted that both the attributes of confidentiality and privacy, though they emerge as implications of errors in *IS*, may not be central requirements for *all* information systems, as there can be information where confidentiality and privacy may not be required. In other words, attributes of security, meaning confidentiality, and privacy may be seen as optional to an information system and depend on the specific context and nature of use of information. Within this frame of reference then, security, privacy can be seen to present extrinsic or subjective attributes of Information Integrity. As indicated earlier, there can be other extrinsic Information Integrity attributes, too, such as usability, independence, sufficiency, understandability, freedom from bias, brief, faithfulness, non-corruptibility, credibility, concurrency, etc. In the Information Integrity research, this categorization of information attributes is pregnant with the development of *Usefulness-Usability-Integrity* paradigm.

6. INFORMATION INTEGRITY RESEARCH & EDUCATION – A GENERAL TOPOGRAPHY

This visualization then focuses the research and education issues in Information Integrity to the study of Accuracy, Consistency and Reliability of information system and of information processed by it. In other words this brings in the research queries of: information origination integrity, storage integrity, retrieval integrity, validation integrity, processing integrity, communication and distribution integrity, *use* integrity, and of information discard or storage for future use integrity. In a broader manner these could be presented as issues of information content integrity, information process integrity and information system integrity: their definitions, measurements, and methods and technologies for their enhancements.

Rajaraman observes that the issue of *Information Integrity* is valid at the data origin stage, at the communication channel stage (comprising medium and/or people), at the processing stage and at the output, i.e., at the information use stage [2]. Thus *Information Integrity* goes beyond the subject matter of data integrity and further covers the requirements of process integrity, medium integrity, people integrity and the output integrity; all these requirements together ensuring the system integrity. As a result, *Information Integrity* emerges as a holistic and fundamental or basic requirement of an *IS* and the information processed by it.

Seen from another angle, with reference to the information system development life cycle, the above then also presents the requirements of ensuring design integrity, development integrity, implementation integrity and maintenance integrity.

7. INFORMATION POLLUTION – ISSUE AKIN TO ENVIRONMENTAL POLLUTION

This is the global nature of *Information Integrity*. In the face of continuous and relentless movement toward system integration maximization so as to reduce the costs of information processing and have competitive advantage, the computerized information systems are becoming complex. This is making them very vulnerable. Vulnerability is further heightened with the widespread use of distributed computers connected by communication networks. All this is working toward lack of integrity in information systems characterizing them (*IS*) with possibilities of incorrect system operation as also massive amounts of polluted (error-filled), i.e., inaccurate, inconsistent and unreliable data and information which are useless and, in situations, even dangerous. This can then be seen as the problem of Information Pollution which if not attended in time will only accelerate the time of *heat death* of the *IS*.

8. SUMMING UP

8.1 Information Integrity Knowledge Development Initiative – A Critical Requirement

In the early days of computing, compared to shared environments of today, uncertainty present in information was minimal. Even then, as mentioned in the beginning, research in data errors received considerable attention. It was experimental research aiming at determining frequency of various types of errors and finding solutions; and it revolved around real data. However, this type of study of science of data errors has long ceased. Later advances in data-related information processing technology mainly pertain to on-line access methods, data transmission accuracy, data security, data encryption, and many other techniques for manipulating and protecting existing data. Emphasis, therefore, shifted from the research and technology for enhancement of data integrity to preserving its statusquo. The assumption is that data is perfect, once validated, and most information processing systems do not anticipate defective data.

And, to the utter discomfort of the direct and indirect users of information system, in the wake of unparalleled growth in system integration technology, this assumption is not sustainable. In other words, computerized, networked information systems of today contain errors that are made but not corrected in spite of controls that are implemented in the logical bounds of applications. The seriousness of the problem at hand becomes further clear when it is recognized that one is dealing with a cybernetic view of an activity whereby ‘data transformation to information’ (which can also be represented as decision processing model) becomes a core *IS* model of an activity. In actual system, these elements of the core model are combined in a variety of ways. They may be repeated, paralleled, and interrelated. Further, output from one set of elements (information) may become input to another set (data) and so on; thereby making the assumption of perfect data wholly untenable and leading to errors in *IS*.

These errors causing information pollution are because, firstly, one is dealing with ‘information’ and not ‘data’, they being different. As a result, one must deal with information errors as against data errors. Information being the ‘data processed’ for ‘use’, as argued earlier, this calls for accounting for process errors and information *use* errors. Secondly, information model that one is now dealing with is at a higher level and has its own brand of uncertainties giving rise to their own types of information errors which must be researched for their integrity implications. And, finally, the *IS* that one is now dealing with is a complex, open system, characterized by complex error mechanisms (not encountered in non-integrated *IS* of yesterdays), which are due to

system environmental uncertainties across the system development and implementation life cycle and come with delay.

Thus, what is called for is study of information errors as above, and research and technology for the enhancement of Information Integrity of information systems and of information therefrom. Current integrity approaches mostly pertaining to data preparation, validation, repair, and assurance are inadequate and ad-hoc, lacking comprehensive methodologies for the purpose. Even critical information processing system logic such as data base updating and information retrieval largely assumes perfect data. In the absence of Information-Integrity-oriented techniques, result reliability, therefore, cannot be guaranteed and defective information will lead to further information contamination – the problem of information pollution. Critical that an *IS* has become to our every day life, solving the information pollution problem as this, complex that it is, offers an opportunity to understand the global nature of *Information Integrity (I*I)* by studying it in new (system's) ways and finding radically new approaches for achieving it. Where it may take information science is to defining what could be termed as *Information Integrity Space (I*I S)*.

When visualized, *I*I Space* could be described as a conceptual aggregation of different dimensions of I*I with many perspectives. All dimensions have one thing in common: they all are related to Accuracy, Consistency, and Reliability of information. The other dimensions of I*I Space are: Data forms: Numeric, Text, Graphics, Voice, Video, Digital, Analogue; Content, Process, System, Informational processes of information origination, storage, retrieval, validation, processing, communication and distribution, *use*, discard or storage for future use; Security, Audit, Control; Prevention, Monitoring, Verification, Detection, Correction; Design, Development, Operation, Use, Maintenance; Manual, Batch, Online, Interactive; Risk Factors; Economic impact; Application Focus; User Domain; Knowledge continuum: Technical, Economic, Social, Ethical, Philosophical; etc.

8.2 Networking for I*I Knowledge Development

Information Integrity Space with its interdisciplinary character possesses potential for new knowledge, science, even an industry. However, to exploit this opportunity will require networking by academia, professionals, practitioners, thought leaders and organizations interested in research, education and communication of clear, compelling, and consistent messages about pervasive and key nature of *Information Integrity*.

REFERENCES

1. Mandke V. V., Nayar M. K. and Malik Kamna, "Information Envelope and its Information Integrity Implications: For a complex, changing environment, modeling a generic business process as an integral to a closed loop information and control system characterized by uncertainty", Proceedings of the 2001 MIT Conference on Information Quality edited by Elizabeth M. Pierce and Raissa Katz-Hass, Cambridge, Massachusetts, USA, November 3-4, 2001.
2. Rajaraman V., "Information Integrity - An Overview", Proceedings of the JNCASR and SERC Discussion Meeting at IISc Campus, Bangalore on Information Integrity - Issues and Approaches, Rajaraman V. and Mandke Vijay V. (Ed.), Bangalore, India (1995).

0-0-0-0-0