Science and Technology Issues Relating to Data Quality in C2 Systems

16th ICCRTS
Paper #031

Jonathan Agre
IDA Information Systems and Technology Division
jagre@ida.org

M. S. Vassiliou
IDA Science & Technology Division

Corinne Kramer
IDA Science & Technology Division

Organization of Talk

- Introduction and definitions
  - Data quality as a lens to examine C2 systems
  - Examples of Data Quality issues in C2
- Comparison of some Data Quality approaches
  - Total Data Quality Management
  - ISO 8000 and 25012
  - DoD and IC
- Two example areas of S&T needs in C2
  - Interoperability
  - Data volume
- Observations and Conclusions

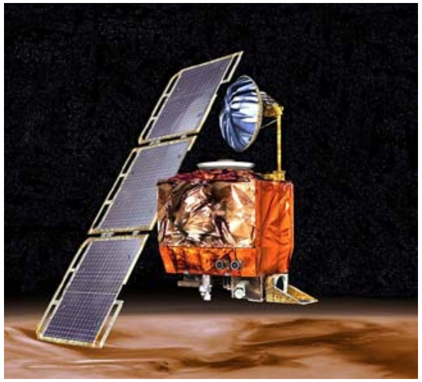
Data Quality

- Bad data results in
  - Errors, mistakes, etc
  - Delays (due to need to reconcile)
  - Loss of credibility
  - Loss of trust
  - Compliance problems
  - Customer dissatisfaction
  - …..many other consequences
- Commercial arena: bad data costs ~$600B annually
- C2 systems operating on bad data can have disastrous consequences in a military setting

Paramount Pictures

Some Well Known Effects of Poor Data Quality

- Mars Climate Orbiter “Metric Mixup”
- Chinese embassy bombing in Belgrade
  - Old map data, not updated with new location
- Challenger Disaster - 1996
  - ...failures in communication... resulted in a decision to launch based on incomplete and sometimes misleading information…
- USS Vincennes downs Iranian Airbus - 1988
  - Lapses in accuracy, completeness, consistency, relevance, fitness for use
- Operation Anaconda
  - Inaccurate and incomplete intelligence data, interoperability problems between allied forces



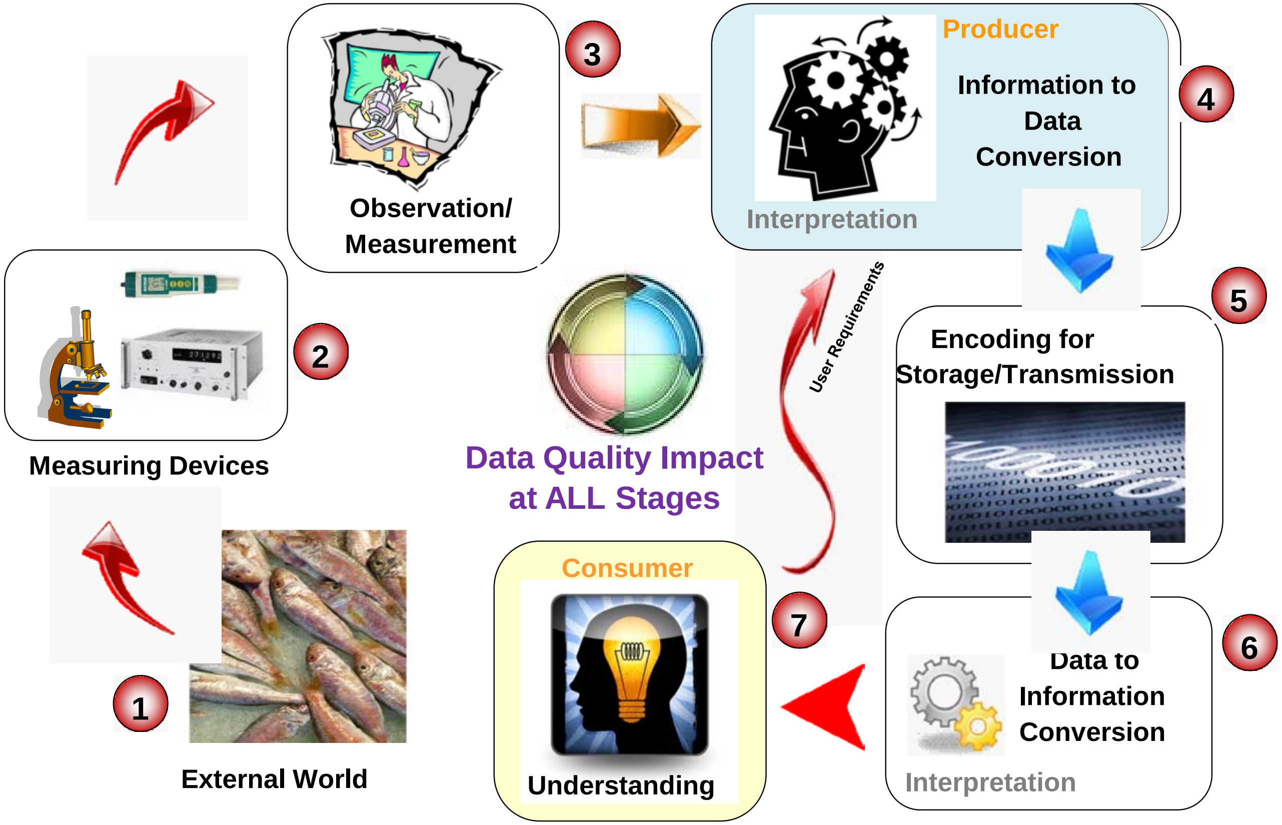
http://nssdc.gsfc.nasa.gov/image/spacecraft/mars98orb.jpg

http://www.wired.com/gamelife/2008/08/rumor-medal-of/

Definitions

- Information: knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning.
- Data: the re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.
- In our context, data includes
  - raw and processed information
  - “all data assets such as system files, databases, documents, official electronic records, images, audio files, Web sites, and data access services.”

Information and Data Production and Consumption Cycle



1

External World

2

Measuring Devices

3

Observation/ Measurement

Producer

Information to Data Conversion

Interpretation

4

5

Encoding for Storage/Transmission

6

Data to Information Conversion

Interpretation

7

Consumer

Understanding

User Requirements

Data Quality Impact at ALL Stages

Loaiza et-al, “Development of a Data Quality Framework for Creating and Maintaining Army Authoritative Data Sources (draft Final),” IDA D-4275, http://mda.ida.org/CIO-G-6-Deliverables/d-4275/D-4275_DraftFinal.pdf

Data in C2

http://militarytimes.com/projects/land_warrior/

- Data used to develop
  - Situation awareness
  - Common operating picture (COP)
- by which commanders make decisions and effect control

- Commanders require many types of data such as:
  - Tactical information
  - Intelligence
  - Logistics
  - Weather
  - Geospatial information

Data in C2

- Data must be collected, analyzed and communicated via various manual and automated messages and exchanged between various C2 systems and people.
- Each C2 system may store portions of the current data and maintains some amount of past data for historical analysis purposes.
- Tempo of activity and volume of data C2 depends on are both rapidly increasing, showing many stress points in the current systems.
- In general terms, a modern C2 system is a large, heterogeneous distributed processing system that is resource limited (bandwidth and computation power) with frequent disruptions and highly dynamic information flows.
- Data are contained in multiple, distributed storage facilities and heterogeneous databases.
- Modern C2 systems, especially in a coalition environment, are among the most complex systems imaginable.

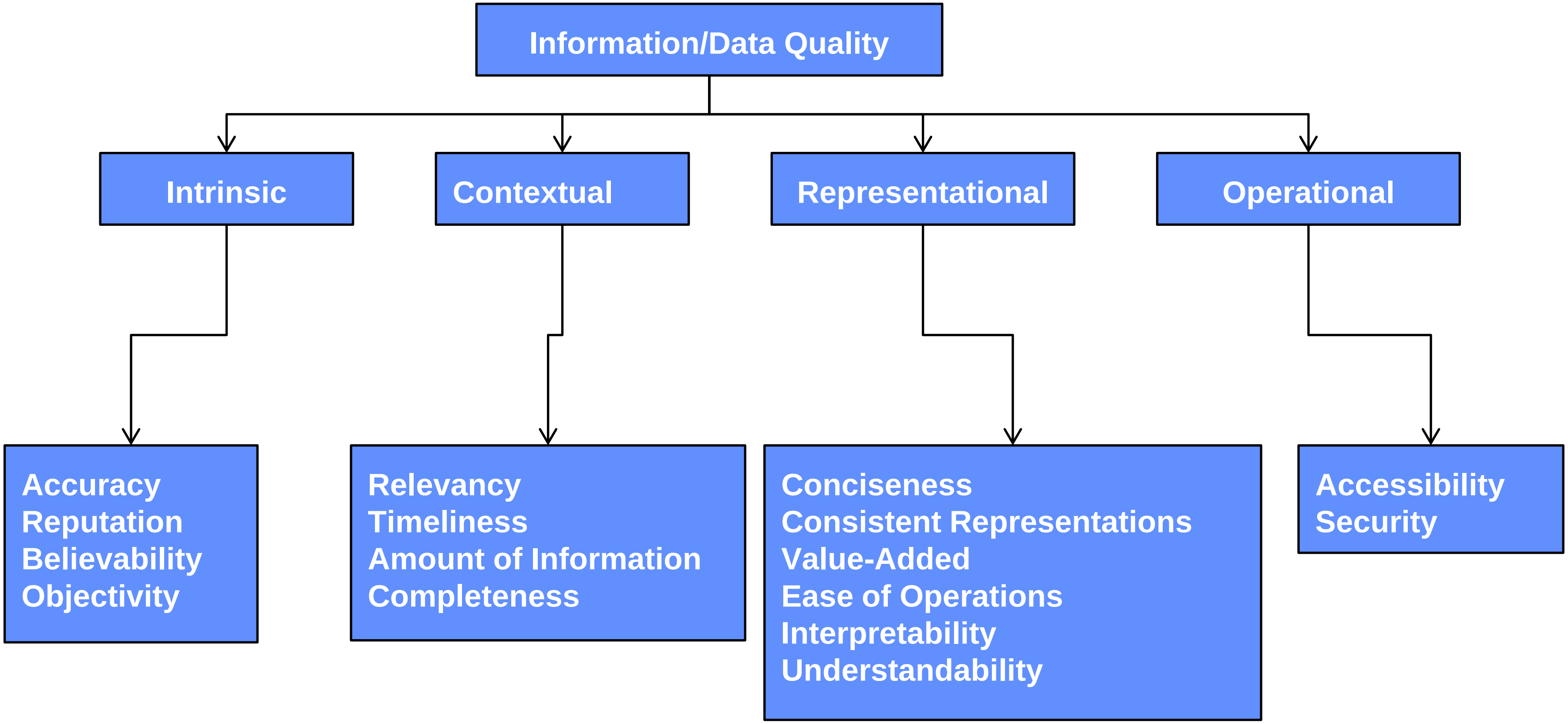
Data Quality

- Information (Data) Quality can be simply defined as "the fitness for use of the information".

- A more practical definition is the degree to which "information and data meets the requirements of its authors, users, and administrators."

- In early data quality research, data was primarily characterized by "ACTS"
  - Accuracy
  - Completeness
  - Timeliness
  - Standards

Total Data Quality Management (TDQM)



Information/Data Quality

Intrinsic

Contextual

Representational

Operational

Accuracy
Reputation
Believability
Objectivity

Relevancy
Timeliness
Amount of Information
Completeness

Conciseness
Consistent Representations
Value-Added
Ease of Operations
Interpretability
Understandability

Accessibility
Security

R. Wang, *Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems*, 1996. 12(4): p. 5-33

ISO 8000 Standard on Data Quality

- Primarily aimed at quality facets of automated information exchange for purchase of goods
- Oriented towards logistics, manufacturing, ERP
- Defines formats for consistent and unambiguous descriptions of individuals, organizations, locations, goods or services.
- Addresses 5 characteristics that define data quality:
    - Syntax
    - Provenance
    - Completeness
    - Accuracy
    - Certification
- and the processes needed to assure data quality

ISO 8000 Standard on Data Quality

- “Part 110: Master Data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification” completed in 2008.
  - Master Data = data held by an organization that describes the entities that are both independent and fundamental for an enterprise, that it needs to reference in order to perform its transactions.
  - e.g. descriptions of customers, suppliers, products, locations, etc.
  - ISO 8000 110 focuses on requirements for exchange of master data that can be checked through automation.

- Representation and exchange of information about provenance (Part 120), accuracy (Part 130), and completeness (Part 140) have also been recently published.

- Other relevant associated standards – 22745 (Open Technical Dictionary)

ISO 25012 Software Engineering Data Quality Model

- Data quality from the software perspective
  - structured data stored in computer systems
- 15 characteristics divided as intrinsic (to the data) and extrinsic
- Some key differences from TQDM and ISO 8000
  - No provenance information
  - Operational notions such as performance, portability, recoverability and availability

Intelligence Community Attributes of Data Quality

- Accuracy
  - Technical errors, misperceptions, deliberate efforts to mislead.
- Objectivity
  - Deliberate distortions and manipulations due to self-interest.
- Usability
  - Compatible with a customer's capabilities and ready when needed.
- Relevance
- Readiness
  - Responsive to the dynamic requirements
- Timeliness

http://levgrossman.com/wp-content/uploads/ 2010/06/spy-vs-spy-without-bombs-775529.jpg

Intelligence Community and C2 Data Quality Issues

- Data sharing and accessibility
  - Much public attention since 9/11
- Spoofing- injection of false data which can corrupt the decision or analysis.
  - Great need to track sources and intermediate handling of data to detect deliberate deception attempts (provenance)
- Inconsistent data that can arise from multiple observers
- Non-authoritative sources of data
  - Often the case, and proper weighting is needed
- In some C2 systems, such as GCCS, the data are generally vetted and considered authoritative, while in others, such as TIGR, the data can be entered by any user that observes an interesting event
  - Both types of systems have their uses; however, provenance should be an explicit factor
  - Information that was presented as true may later be found to be untrue
- Timeliness and accuracy can have a more severe impact in a C2 tactical situation

DoD

- DoD recognizes data quality issues
- Data quality discussed in several key documents:
  - DoD Net-Centric Data Strategy (NCDS), May 2003
  - Data Sharing in a Net-Centric Department of Defense, Dec 2004
  - Guidance for Implementing Net-Centric Data Sharing, Apr 2006
  - DoD Command and Control (C2) Strategic Plan Version 1.0, Dec 2008
  - Interim Guidance to Implement NCDS in the C2 Portfolio, Mar 2009
  - DoD C2 Implementation Plan Version 1.0, Oct 2009.

  - DoD Information Enterprise Strategic Plan (2010-2012)
  - Army Directive 2009-03 Army Data Management
  - Army Knowledge Management and Information Technology AR 25-1 (coming)

DoD Net Centric Data Strategy (NCDS)

Defines 7 goals for the data strategy:

- Visible (who has and what)
- Accessible (where and what format)
- Understandable (what is meaning)
- Institutionalized (what and who governs)
- Trusted (trustworthy, accurate and authoritative)
- Interoperable
- Responsive

Army Data Transformation (ADT)

Working to improve data quality in 6 dimensions:

- Accountable
  - incorporate common data standards and governance practices.
- Authoritative
- Transform
- Expose
  - Messaging, Data Services, Data Warehouses and Data Security
- Register (e.g., authoritative data repository)
- Assess

Army Data Services Layer

- Provides application services for standardized handling of data
- Part of the Enterprise Information Architecture

Cross-Mapping of Data Quality Concepts

TDQM	DoD NCDS Data Goals	Intelligence Community	ISO 8000	ISO 25012
Intrinsic:				
Free-of-error	Trusted	Accuracy	Accuracy	Accuracy, Precision
Reputation	Trusted (Accountable Authoritative)		Certification	
Believability	(Accountable Authoritative)		Certification	Credibility
Objectivity (Provenance)	Trusted (Accountable Authoritative),	Objectivity	Provenance	Traceability
Operational (Accessibility):				
Accessibility	Visible, Accessible (Expose)	Usability		Accessibility, Availability, Portability, Recoverability Performance
Security (Access Control)	Trusted (Expose)			Confidentiality
Contextual:				
Amount of Information				
Relevance	Responsive to Users Needs	Relevance, Readiness		
Value added				
Timeliness	Responsive to Users Needs	Timeliness		Currentness
Completeness			Completeness	Completeness
Representational:				
Understandability	Understandable	Usability	Master Data encoding, Open Tech. Dict.	Understandability
Conciseness				
Ease of operation				Performance
Interpretability	Interoperable	Usability	Master Data Syntax	
Consistent Representations	Institutionalized, Interoperable (Transform, Register)		Master Data: Conformance	Consistency. Compliance

Comparison

- Intelligence community "usability" covers several areas and would be difficult to measure
- TDQM does not capture the notion of readiness (data adaptable to changing circumstances and requirements)
  - Not explicit about governance
- ISO 8000 provides a broad range of coverage, but does not address
  - Timeliness
  - Ease of operation
- NCDS does not address several properties critical to C2
  - Particularly timeliness
  - Coverage of believability and reputation primarily limited to using authoritative data sources
  - Does not cover many situations frequently encountered in C2
    - e.g. data from a variety of sources with varying degrees of pedigree (provenance, reliability, etc).

http://www.signandtrade.com/nba/nbafantasytools.aspx

Metrics, Tools, and Processes

- Basic metrics developed for each TQDM dimension
  - ratios, min/max, weighted averages
    - e.g., accuracy = # accurate records/total number of records
  - Some are subjective and difficult to quantify (e.g., believability)
  - Utility is function of application context (weightings)
- Tools (active commercial market)
  - Data validation
  - Extract-Transform-Load
  - Data profiling/data auditing/data cleansing
  - Data Monitoring
- Maintaining data quality
  - Data deteriorates over time
    - e.g., people – death, marriage, divorce, change of address, etc
  - Information Production Maps
    - Data is a product that goes through “manufacturing” stages

C2 Critical Data and System Issues

- Interoperability
- Information Overload (Volume)
- Communications Limitations
  - Degraded, Intermittent, Limited Bandwidth
- Distributed Access
- Timeliness
  - Including realtime processing for inclusion in COP
- Accuracy
- Provenance
- Security

Interoperability

- Each service, each coalition partner, has its own family of C2 systems
  - GCCS FoS comprises over 200 systems or services
  - Data must be exchanged between different systems, families

- Universal Core (uCore), currently version 2.0
  - Information exchange specification and implementation profile that defines a vocabulary of commonly exchanged concepts such as who, what, when and where
    - Further developments: temporal relationships; allowing items to be of different types at different times
  - Joint Consultation, Command and Control Information Exchange Data Model (JC3IEDM) used by NATO.
  - C2Core: DoD high-level data model for C2 based on JC3IEDM; extension of uCore
    - UCore requires extensions to include the full JC3IEDM and that there will still need to be a mapping of JC3IEM to C2Core

Interoperability (Cont.)

- Key S&T issue relating to interoperability:
  - Automated methods to resolve differences among the semantics of the differing systems
  - Even with standardized data exchange methods, there will be subtle interpretations of data that will need to be resolved by human intervention
  - There are too many relationships among data for people to represent and capture all of the relations between the entities involved and the overall process would benefit from automation
    - Reasoning over unstructured data

Data Volume

- Explosion of raw and processed data entering C2 systems
  - Continued growth expected
  - More sensors, video, UAVs, etc
- Large volumes of data with poor control of data quality

http://ourmastermindsgroup.com/blog/wp-content/uploads/2011/03/information-overload1.jpg

Data Volume (Cont.)

http://ourmastermindsgroup.com/blog/wp-content/uploads/2011/03/information-overload1.jpg

- Key S&T challenges in handling large volumes of data:
  - Processing architectures to analyze the data
  - Methods for securely sharing data and results
  - Management of large data sets, including multilevel classifications
  - Disadvantaged communications
  - Decision support
    - At least 26 research projects in 2009
  - Data-to-decision program to address some of these issues
- Some analysts have suggested greater emphasis should be put on assisting users to understand the information rather than designing for full automation
  - In either case, additional emphasis should be given to understanding the data quality and incorporating this into the decision processes.

S&T Research Issues

- Much research on general aspects of data quality, but little directed towa C2
    - e.g., Automated provenance handling
        - But still very difficult to determine if a document has been copied or combined, unless it has been under version control for its entire existence.
        - “Ringing” problem in intel/situation reports
    - Need to consider the various types of C2 data when considering how to capture C2 quality features.
        - e.g., Quality features of raw data may be very different from a command message or a situation report
    - Can one provide appropriate metadata along with every data item so that the data becomes self-describing and self-protecting?
        - How best to accomplish this within the constraints of limited bandwidth, processing power, intermittent service in a disruption-tolerant and robust fashion?
- Methods required to reduce the load on the commander through automatic processing
    - Semantic processing of structured and unstructured text

Next Steps

- Examine several specific C2 systems for data quality characterization and requirements
- Define and develop a set of C2-specifc data quality dimensions, metrics, associated weightings and quality tools
- Incorporate current standards, such as ISO 8000 and 25012, tailored for C2 applications, into C2 systems.
- Continue to set policies and governance to accomplish data quality goals and methods of enforcement.
- Define standard data and metadata services. The ADSL is an excellent start at this process.
- Data quality characteristics and their associated metrics should be explicitly incorporated into C2 systems and processes
  - Include some forms of provenance information with the data items in decision support systems
- Interoperability and data sharing should continue to be addressed by the C2 community including rapid methods of modifying and updating the data exchange standards

Conclusions

- TDQM: 16 general characteristics of data quality have broad community acceptance
- C2 systems are beset with similar data quality issues as is the general enterprise IT community.
- All the TQDM data quality characteristics are relevant to C2 systems.
  - However, several quality issues are of relatively greater importance to C2 due to potential lethality of errors in decision making.
- Quality characteristics are not independent and should not be addressed in isolation, but should be part of an ongoing process, including governance.

Enabling future C2 systems to explicitly and automatically incorporate data quality dimensions of the underlying data will improve the decision making ability of the commanders by reducing the uncertainty in their decision space

Backup

Data Quality: TDQM

- Total Data Quality Management (TDQM) research community has expanded ACTS to 16 quality dimensions:

Intrinsic Properties

- Accuracy/Freedom-from-Error - The extent to which data is correct and reliable.
- Believability - The extent to which data is regarded as true and credible.
- Reputation -- The extent to which information is highly regarded in terms of its source or content.
- Objectivity - The extent to which data is unbiased, unprejudiced, and impartial.

Accessibility Properties

- Accessibility -- The extent to which data is available, or easily and quickly retrievable.
- Security - The extent to which data access to data is restricted appropriately to maintain its security.

Contextual Properties

- Relevance - The extent to which data is applicable and helpful for the task at hand.
- Timeliness - The extent to which data is sufficiently up-to-date for the task at hand.
- Completeness - The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.
- Amount of Information - The extent to which the volume of data is appropriate for the task at hand.
- Value Added - The extent to which data is beneficial and provides advantages from its use.

Representational Properties

- Conciseness - The extent to which data is .compactly represented.
- Consistent Representation - The extent to which the data is presented in the same format.
- Ease of Operations - The extent to which data is easy to operate on and apply to different tasks.
- Interpretability - The extent to which data is in appropriate languages, symbols, and units and the definitions are clear.
- Understandability - The extent to which data is easily comprehended.