

A Model for Data Quality Assessment

Baba Piprani¹, Denise Ernst²

¹ SICOM Canada

² DSI Now, Canada

babap@attglobal.net, denise.mcconnell@rogers.com

Abstract. One of the major causes for the failure of information systems to deliver can be attributed to data quality. Gartner's figures and other similar studies show the failure rate hovering at a plateau of 50% for data warehouses since 2004. While the true cause of poor data quality can be attributed to a lack of supporting business processes, insufficient analysis techniques, along with protecting oneself with the introduction of data quality firewalls for incoming data, the question has to be raised as to whether a data quality assessment of the existing data would be worthwhile or plausible? This paper defines a data quality assessment model that enables a methodology to assess data quality and assign ratings using a score-card approach. A by-product of this model helps establish 'sluice gate' parameters to allow data to pass through data quality filters and data quality firewalls.

Keywords: Data Quality Assessment, Data Quality Firewall, Data Quality filter, Data lineage, Type Instance

1 Introduction

Did you know that in September 1999 a metric mishap caused the crash landing of a Mars bound spacecraft where NASA lost a \$125 million Mars orbiter because 2 engineering teams, for a key spacecraft operation, used different units of measure which resulted in failure of data transfer due to the mismatch?[2]. Did you know that a referee in the World Cup Soccer 2006 match between Australia and Croatia handed a soccer player 3 yellow cards before the player was sent off? The rule is 2 yellows results in a player being sent off [1]. Did you know that data warehouse success measures, or more appropriately stated, "failure rates or limited acceptance rates" have been in the range of 92% (back in late 1990s) to greater than 50% for 2007 [3][6]---a dismal record indeed.

So what do we mean by "failure"? The meaning of the term "failure" has been amplified by the Standish Group [4] with the interpretation that the "success" of the project refers to the project being completed on time and on budget with all features and functions as initially specified; or, the project being "challenged" refers to the project is completed and operational but, over-budget, over the time estimate, and offers a subset of features and functions originally specified; and being "impaired" refers to the project being cancelled at some point during the development cycle.

According to the Standish Group's 2003 CHAOS report, 15% of the IT projects "failed" and another 51% were considered "challenged", while 82% of the IT projects experienced significant schedule slippage with only 52% of required features and functions being delivered. For 2004, results show that 29% of all projects succeeded i.e delivered on time, on budget, with required features and functions; 53% were "challenged"; and 18% failed i.e. cancelled prior to completion or delivered and never used. A staggering 66% of IT projects proved unsuccessful in some measure, whether they fail completely, exceed their allotted budget, aren't completed according to schedule or are rolled out with fewer features and functions than promised [5].

2 Root cause of failures

Quality appears missing in the meaning of success or failure of a IT project. Lack of data quality appears to be the major culprit in the "failure" of IT projects. The traditional project management triangle is represented by cost, scope and schedule in Figure 1 with quality injecting throughout the cycle but often enough there no associated project deliverable. This gap is unexpected yet understood so how do we assess and close the gap?

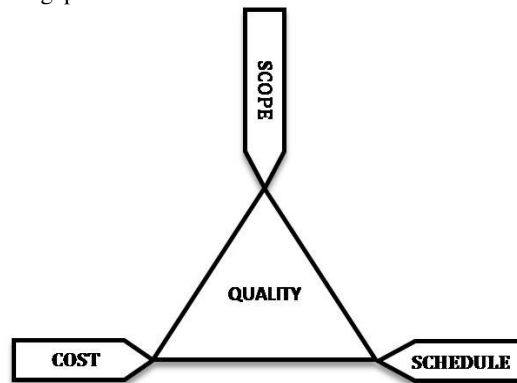


Fig. 1. Project Management triangle.

It is important to observe that in any typical manufacturing process, quality is injected in every process from the very start. For example, a casting metal foundry technician systematically monitors a melt to assess that the required composition of metal compounds like carbon, iron, nickel, chromium, zinc etc. are in place prior to pouring to ensure the quality of the desired metal casting. Similarly, it is imperative that data quality needs to be injected in every phase of information system design and implementation with due diligence to governance, monitoring and auditing, among other things. In this paper we explore how we can define a similar quality control assessment for data.

3 Issues to tackle

Where do we start? In our experience, examining the IT projects we have been called into to salvage and steer the usually sinking project towards 100% success, we have observed the following issues:

- Business requirements documentation is non-existent, not maintained upon change, too high-level, lacks integrated enterprise viewpoint, and lacks supporting business processes
- Business rules are buried in program code which results in higher maintenance costs, dependency on specialized skills, and a lack of awareness
- Undocumented definitions and missing semantics
- Inability to audit and monitor changes to the architecture and contained data

These are only samplings of the issues that are encountered that contribute to the data quality chasm in the building of information systems in both existing and under development. This list demonstrates a need to address data quality assessments throughout the solution's Systems Development Lifecycle (SDLC) as in Figure 2.

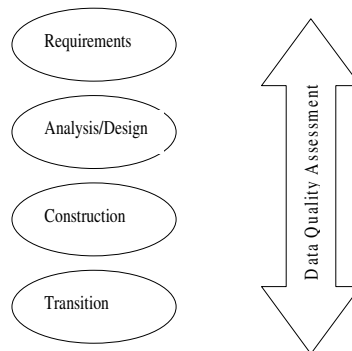


Fig. 2. Generic SDLC and data quality assessments.

4 Data Quality Assessment Objective

The objective of the assessment is to identify the quality of the data in the identified business activity. An organization could be primarily in the service industry while another in the regulation sector.

The assessment results determine the accuracy, completeness, consistency, precision, reliability, temporal reliability, uniqueness and validity of the data. Assessment standard criteria are used when conducting the assessment.

When conducting an assessment, the business requirements are your window to the world. This can be a daunting task when assessing the quality of data at the enterprise perspective. Remember to scope the tests within the assessment criteria to ensure a balanced cost/benefit for the organization.

For example, if financial data is the vital to the business operations then the quality of such data will be an important factor in key business decision-making. Quality Assessment can happen in several manners, generally as either detection tests or penetration tests.

- Assessment detection tests assess data quality, identify risks, and can be used to determine risk mitigation efforts.
- Assessment penetrations tests, in addition to assessment detection, will penetrate systems with faulty data and monitor the effect and result. This will help to identify process deficiencies as well as determine quality of the data.

The summary of assessment tests should reveal data quality scorecard metrics vital to the organization's business and operations.

5 Methodology

Assessing data quality should not be like trying to pick up jello! Nor should it be an exercise in throwing darts on a Saturday afternoon in a pub! What is needed is an approach to methodically put in place data quality measures and standards sufficiently applicable at any stage of the life cycle, even if being parachuted in any part of the life cycle!

Then it becomes a primary requirement to be able to assess data quality across both the earlier stages and later stages of the development life cycle from any given point in the development life cycle. Not only that, it should be possible to precisely home in on any given stage of the development life cycle to enable the establishment of subsequent correctional measures going forwards.

Table 1 highlights how data quality assessment criteria are addressed by NIAM and ORM based modeling. The data quality assessment tests that can be conducted in level pair constructs across the 3 data lineage levels to determine the resultant data quality. The assessment test examples can be performed based on the data lineage level of the attribute e.g. by allocating each attribute in the implementation with a 'class-term' which simply groups similar attribute types based on similarity of concepts, e.g. amounts, dates, ratios, counts, quantities.

To achieve an organized approach for assessing data quality, the phases of a generic systems development life cycle could be aligned into 3 data lineage levels:

- Terminology and semantics
- "Type" - (metadata)
- "Instance" - (value)

The data lineage level pairs can be assessed in 'type – instance' level pairs. The 2 level pairs are "[terminology and semantics) + (Type metadata)]" and [(Type metadata) + (instance value)]"

The following criteria were found to be helpful in assessing data quality at each data lineage level as applicable (no ordering priority, alphabetical):

Table 1. Data Quality Assessment Criteria.

Assessment Criteria	Description	Applying NIAM / ORM	Schema data assessment test
Accuracy	Degree of agreement between a set of data values and a corresponding set of correct values i.e. is the data correct	Natural language sentences with accompanying population diagrams and realistic sample data value populations related in a fact type that associated with a business “concept”	Where numeric data like amounts, counts, and quantities are involved, look for min and max ranges, less than or greater than zero or other figure etc. Where dates are involved look for future, past checks. Look for all of the percentages adding up to 100 or some defined limit.
Completeness	Degree to which values are present in the attributes that require them i.e. is the data complete	Null ability vs. totality as applied to a concept based fact type along with incomplete sentence populations in a population diagram.	Look for name required for regulated item records e.g. a NULL cannot own an aircraft If a large value property mortgage has been declared as being owned by a group, then the percent of the group ownership must add up to 100%
Consistency	Agreement or logical coherence in accordance with facts without variation or contradiction,	Using a “business concept” based focus; natural language sentences with sample real data values form the basis of agreement and understanding with the business. Expanding the natural language sentence constructs with variations and contradictions in population that break business rules.	Look for consistency in dates; like you cannot die in the future, you cannot be born in the future.
Precision	The quality or state of being exact, within the defined bounds and goals	The natural language elementary sentences and in particular, the elementary sentences form the basis of exact precision of the statement of the business fact in unambiguous terms	Look for precision in location latitude and longitude declarations; they must contain seconds in the Degree/Minute/Second system. Look for rounding errors
Reliability	Agreement or logical	Using a “business concept” based focus,	Look for consistency of business types that an

	coherence that permits rational correlation in comparison with other similar or like data	natural language sentences with real data values form the basis of logical coherence amongst the users to identify pattern similarities, e.g. risk measures used for different insurance companies based on insurance company groupings can be identified as being common across the company.	organization is licensed for and related types of returns or transactional consistencies Look for lack of referential integrity on the use of same attributes being used in various tables
Temporal Relatability	Meanings and semantics that can change over time	Adding time dimension to the natural language sentences where the same fact type may undergo transformations.	Applicable to uniquely traceable items like serial numbers or particular licensed item identifiers, look for can the same item be involved with another item at the same time. Applies to ownership, involvement, and lineage. Look for loss of history data with no record of previous values Look for errors in versioning
Timeliness	Data item or multiple items that are provided at the time required or specified	The granular level of the NIAM and ORM models provide easy cross correlation with data sets to be examined for timeliness in terms of availability and for the derivation of sequencing requirements thus directly affecting audit control.	Data extractions Transactional details, extractions and loads to be monitored for same-period reliability Inability to relate data due to loss of history data
Uniqueness	Data values that are constrained to a set of distinct entries—each value being the only one of its kind	Every fact type must have at least one uniqueness constraint. This aspect is built into every sentence type and population diagram example for each concept and fact in NIAM and ORM	Look for dangling tables. Look for lack of referential integrity on the use of same attributes being used in various tables
Validity	Conformance of data values that	The lower level granularity of a fact type	Look for artificial keys, identity values, system

are edited for associated with a business generated keys and apply at
acceptability— concept and the least one business key to a
reducing the population diagrams helps data grouping say in a data
probability of define the acceptability mart or row occurrence for
error and validity of data. a registry type data group
(an inventory list like list of
persons, list of vehicles etc)

5.1 Terminology and Semantics Data Quality

Taking advantage of the disciplined process for deriving business semantics and business rules in ORM, we see how a terminology based approach affects the basic data quality characteristics that can be used for assessing data quality.

In other words, use of SBVR with NIAM and ORM can contribute extensively to mitigating the risks in the approach to data quality assurance and definition of an enterprise model in the discovery of ‘missing business rules’.

5.2 “Type” (Metadata) Data Quality

Ideally, a well defined approach to assess data quality would be to first properly defining business requirements using say, SBVR using common terminology [8] and then following the bridging from the business model to an information system involving metadata and data models, and transforms for Terminology to Semantic Metadata and data models for:

- Business ontology to consolidated data and rules requirements
- Business requirements to class of platform independent data models
- Class of platform independent data models to class of platform specific data models
- Class of platform specific data models to vendor platform specific data models.

However, in reality it is a backstitched approach that works i.e. reverse engineer to look for ‘hidden business rules’ using some form of fact modeling methodology.

We see metadata here as being applicable to the result of a transform from terminology to data models---be they platform independent data models, platform specific data models or vendor platform specific data models.

The same above data quality characteristics assessment criteria can be applied to metadata at the level where a metadata registry-of-sorts e.g. ISO 11179:2003 Information technology — Metadata Registries (MDR) Part 3: Registry Metamodel and Basic Attributes [11]---is defined or available for an organization. Data quality assessment at the metadata level can now be performed with respect to other metadata elements in the registry in a similar fashion in correlation with other metadata elements---of course, with the similar transforms---being able to compare like metadata, particularly when different representations are involved with the same object type [9].

5.3 Instance (“Value”) Data Quality

One of the aspects that would be helpful in the assessment of the next level for “Instance” Data Quality (“value”) is the ability to match like data elements that form the metadata, i.e. a match between a “Customer_number” with “Customer_id”--- which could provide different answers [9]. An important aspect in this regard is the use of standardized “class terms” for the data element names or categories. By identifying a data element e.g. an SQL column with meaningful data element names associated with corresponding constraints would also be part of the assessment, e.g. START_DTE, STOP_DTE; or AIRCRAFT_WEIGHT_NBR, PLATINUM_LOW_BALANCE_AMT etc. would also assist in establishing a data quality assessment criteria rules-set, where column suffixes _DTE, _NBR, _AMT are class terms applicable to SQL columns.

To correlate: A class term of column suffix _DTE could be associated with a requirement to be able to validate a range of dates, e.g. STOP_DTE must be greater than START_DTE. Additionally, a STOP_DTE could be defined as not being in the past and always must be greater than the CURRENT_DATE or be in the future. Similarly, the START_DTE must always be in the past or be the same as the CURRENT_DATE and never be in the future. Similarly, a PLATINUM_LOW_BALANCE_AMT for a platinum account cannot fall below \$1000 etc.

These are examples of how syntax based class terms for a column can be used to automatically derive constraining parameters, as imported and derived from the transforms of the SBVR NIAM or ORM based model.

6 Data Quality Metrics

The results of the Data Quality Assessment can provide operational metrics that can be used to continually monitor the quality of data. Where does one find metrics? An example of a simple metric could be the extent of implemented business rules vs. the total set of business rules (which can be determined via analysis using the Semantics and Terminology data lineage level).

In order to establish a metric of this nature, it is necessary to be able to derive the existing and missing business rules from a common template which should apply consistently no matter which level you are at---the instance (value) level, the type (metadata) level or the terminology or semantic level. In other words, you are looking for the same rule being interpreted across data lineage levels towards implementation.

An optimal method to achieve the set of “missing business rules” is to reverse engineer the semantics from the ‘instance’ level, or even at the type level. For example, given a set of values for aircraft mark identifiers (the visible letters identifying any aircraft), and a set of owners of the aircraft, we can look for things like: is it mandatory that an aircraft being registered needs to be owned by a party that has an address and contact, or can the same serial number engine exist on more than one aircraft at the same time, or can the same aircraft mark be on more than one aircraft at the same time etc. This question set can only be answered by conducting a

quick reverse engineering or backstitching exercise from the attribute driven (ER, UML) relational or object based model to a fact based semantic modeling technique (ORM, NIAM, COGNIAM) to discover all hidden business rules. Some proponents might balk at the reverse engineering exercise, but if one examines the instance and type level paired combinations as part of a fact based sentence type, then it is a simple matter to validate that fact type!

The defined metrics can drive a dashboard-style data quality assessment scorecard. A scorecard could be in terms of reliability based on how the data has been preserved over the life cycle since inception, based on say, continuity or commitment to establish data currency by organizational personnel.

For example, a metric for particular data element, which, after going through the 3 data lineage levels in data quality assessment, could be assigned a value of “Yellow”, while another data element could be assigned “Red”---meaning the “Yellow” tagged element is to be relied upon with caution (again, this can be further defined with a breakdown as to if it is coming from the Western Region operations it is 60% reliable vs. Eastern Region operations with only 40% reliable , while the “Red” tagged element could simply state that this data element is not reliable.

Some of the additional activities [10] that support a scorecard include:

- Defining acceptable parameters and tolerances
- The metrics model that includes high risk situations based on data quality tolerances being met, establishing risk mitigating measures and providing definitive assurances vs. hand-waving or ‘I think...’ scenarios
- Go / no-go thresholds
- a metrics evaluation model
- process model to feed metric values into the scorecard

What this is demonstrating is that based on an assessment methodology using Semantics, Metadata, and Instance, one is able to derive a dashboard equivalent value for a given metric that is usable in the scenario for reliability and dependability of the quality of data.

7 Author’s experiences

In the author’s experience over the past 2 decades, assessing data quality has provided significant return on investments in varying private sector industries as well in the public sector.

The data quality assessment methodology as outlined in this paper has been used to recover seriously failing IT development / data warehousing projects; as a basis for Master Data Management; common object mappings across heterogeneous applications; in data warehousing; creation of metadata repositories; data quality firewalls; define enterprise data standards; establishing data quality scorecard; and the list goes on.

The approach has uncovered exceptional findings such as: Non-existing financial coding for expenses charged for over \$120 million; \$150 million recorded as amount already spent in the year 2097; Out of 26000 private owner records for a ‘regulated licensed item’, nearly 16000 were rejected due to inconsistencies like null address,

lack of a proper or complete address, null name, names like Theodore and Alvin Chipmunk, Mickey Mouse etc.; In a mid-sized organization, there are over 100 staff members in an organization who have a birth date of 2061; 100% of HR personnel data was rejected when assessed against ~1000+ 'business rules'; Entities reporting on the number of hours flown in a given year as 10000 hrs. (Note the maximum hours in a year is 365 days x 24 hrs = 8760 hrs).

A few major success stories include, all of which were on time and under budget:

A large financial data warehouse with over 200 entities designed with 100% data quality at the schema level including auditing and monitoring tools; It took a remarkable 90 days with a 8 person team consisting of subject matter experts and technical staff.; The warehouse continues to be operational and has never once been re-loaded; Operational maintenance consists of 2 technical staff and 1 business analyst supported by the infrastructure-working group; Over 100 users accessing 3 cubes and 50+reports; Implementation of a previously failed re-design of a financial system used in the regulatory sector; Resurrected, assessed, data converted and operational within 2 months with a 5 person combined business/technical team ; Data quality scorecard used by management.

References

1. BBC Sport:Ref Poll sent home from World Cup. In http://news.bbc.co.uk/sport2/hi/football/world_cup_2006/5108722.stm
2. CNN.Com:Metric mishap caused loss of NASA orbiter. In: <http://www.cnn.com/TECH/space/9909/30/mars.metric.02/>
3. Gartner Group Report: Gartner Press Release, Gartner Website – Media relations(2005). In: http://www.gartner.com/press_releases/pr2005.html
4. Standish Group International, Inc.:Chaos Chronicles and Standish Group Report. (2003). In: http://www.standishgroup.com/sample_research/index.php
5. Standish CHAOS Chronicles: Lessons From History. In: <http://lessons-from-history.com/Level%20Project%20Success%20or%20Failure.html>
6. ITtoolbox Blogs, Madsen, Mark: A 50% Data Warehouse Failure Rate is Nothing New. In: <http://blogs.ittoolbox.com/eai/rationality/archives/a-50-data-warehouse-failure-rate-is-nothing-4669>
7. John A. Zachman: A Framework for Information Systems Architecture. In: IBM Systems Journal, vol. 26, no. 3, IBM Publication G321-5298(1987)
8. Donald Chapin, John Hall, Sjr Nijssen, and Baba Piprani: A Common Terminology, Semantic Metadata & Data Model Framework for Relating SBVR, ISO 704 & 1087 to ISO/IEC 19763 & 11179. In: Metadata Open Forum, Sydney(2008). see <http://metadataopenforum.org/index.php?id=34,132,0,0,1,0>
9. Baba Piprani: Using ORM in an Ontology Based Approach for a Common Mapping Across Heterogeneous Applications. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2007 Workshops. LNCS, vol. 4805, Springer, Heidelberg (2007).
10. Baba Piprani: Using ORM Based Models as a Foundation for a Data Quality Firewall in an Advanced Generation Data Warehouse. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4278, Springer, Heidelberg (2006).
11. ISO 11179: Information technology — Metadata Registries (MDR) Part 3: Registry Metamodel and Basic Attributes, International Standards Organization, Geneva.(2003)