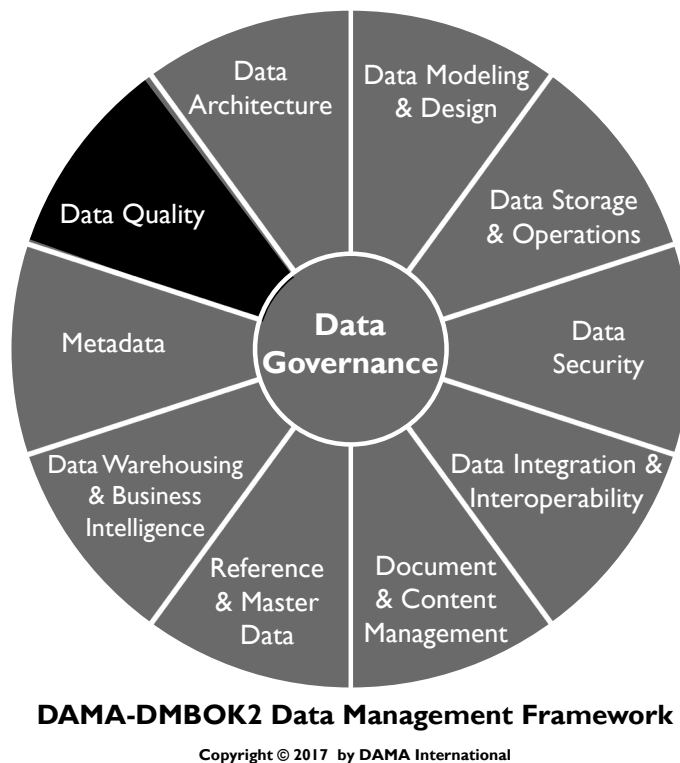


Data Quality Management



1. Introduction

Data Quality can be defined as the degree to which dimensions of Data Quality meet the requirements. This implies that requirements should be formulated for each (relevant) dimension. A much shorter definition for quality of data is 'fit for purpose.'

Data that meets the requirements are of sufficient quality; data that doesn't meet the requirements are of insufficient quality. To keep it simple, we respectively speak of high and low, or poor quality data.

Effective Data Management involves a set of interrelated processes enabling an organization to use its data to achieve strategic goals. An underlying assertion is that the data itself is of high quality. Data Quality

Management is the planning, implementation, and control of activities that apply quality management techniques to data in order to assure it is fit for consumption and meets the needs of data consumers.

High quality data is context driven. This means that the same data may be simultaneously viewed as high quality by some areas of an organization while being viewed as low quality by other areas. Many organizations fail to engage with this question of context, that is, high Data Quality being that which is fit for purpose.

If we understand organizations as data manufacturing machines, we can assert (from our experience in manufacturing) that organizations that formally manage the quality of data will be more effective, more efficient and deliver a better experience than those that leave Data Quality to chance. However, no organization has perfect business processes, technical processes, or data management practices. In reality, all organizations experience problems related to their Data Quality. Many factors undermine quality data: lack of understanding about the effects on organizational success, leadership that does not value Data Quality, poor planning, ‘siloes’ system design, inconsistent development processes, incomplete documentation, a lack of standards, or a lack of Data Governance.

As is the case with Data Governance and with Data Management as a whole, Data Quality Management is a function, not a program or project. This is because projects and even programs have starts, middles, and ends. A Data Quality Function is, or should be, a continuing business as usual set of activities. It will include both projects and programs (to address specific Data Quality improvements) as well as operational work, along with a commitment to communications and training. Most importantly, the long-term success of a Data Quality improvement program depends on getting an organization to change its culture and adopt a quality mindset. As stated in *The Leader’s Data Manifesto*, “fundamental, lasting change requires committed leadership and involvement from people at all levels in an organization.” People who use data to do their jobs – which in most organizations is a very large percentage of employees – need to drive change. One of the most critical changes to focus on is how their organizations manage and improve the quality of their data.⁷¹

Formal Data Quality Management is similar to continuous quality management for other product manufacturing. It includes managing data through its lifecycle by setting standards, building quality into the processes that create, transform, and store data, and measuring data against standards. Managing data to this level usually requires a Data Quality Function team. The Data Quality Function team is responsible for engaging both business and technical data management professionals and driving the work of applying quality management techniques to data to ensure that data is fit for consumption for a variety of purposes. The team will likely be involved with a series of projects through which they can establish processes and best practices while addressing high priority data issues.

⁷¹ For the full text of *The Leader’s Data Manifesto*, see <http://bit.ly/2sQhcy7>.

Data Quality Management

Definition: The planning, implementation, and control of activities that apply techniques for collecting and handling data ensuring it addresses the needs of the enterprise and local consumer and is fit for use.

Goals:

1. Develop a governed approach to make data fit for purpose based on data consumers' requirements.
2. Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
3. Define and implement processes to measure, monitor, and report on data quality levels.
4. Identify and advocate for opportunities to improve the quality of data, through process and system improvements.

Business Drivers



Inputs:

- Data Policies and Standards
- Data Quality Expectations
- Business Requirements
- Business Rules
- Data Requirements
- Business Metadata
- Technical Metadata
- Operational Metadata
- Data Sources and Data Stores
- Data Lineage
- Data Issues
- Data Sharing Agreements

Activities:

1. Define a Data Quality Framework (P)
2. Define High Quality Data (P)
3. Identify Dimensions and supporting Business Rules (P)
4. Perform an Initial Data Quality Assessment (P)
5. Identify and Prioritize Improvements (P)
6. Define Goals for Data Quality Improvement (P)
7. Develop and Deploy Data Quality Operations (D,O,C)
 1. Manage Data Quality Rules
 2. Measure and Monitor Data Quality
 3. Develop Procedures for Managing Data Issues
 4. Establish Data Quality Service Level Agreements
 5. Data Quality Response

Deliverables:

- Data Quality Strategy & Framework
- Data Quality Improvement Program
- Analyses from Data Profiling
- Recommendations based on Root Cause Analysis of Issues
- Data Quality Procedures
- Data Quality Reports
- Data Quality Governance Reports
- Data Quality Service Level Agreements
- Data Quality Policies and Guidelines
- Remediated Data
- Certified Data

Suppliers:

- Business Management
- Subject Matter Experts
- Data Analysts
- Data Architects
- Data Modelers
- System Specialists
- Data Stewards
- Business Process Analysts
- Data Vendors

Participants:

- CDO
- Data Quality Analysts
- Data Stewards
- Data Owners
- Data Analysts
- Subject Matter Experts
- Database and Integration Specialists
- Data Professionals
- DQ Managers
- IT Operations

Consumers:

- Knowledge Workers
- Data Stewards
- Data Professionals
- IT Operations
- Data Governance Bodies
- Partner Organizations
- Centers of Excellence

Technical Drivers



Techniques:

- Data Quality Metrics
- Profiling
- Preventive and Corrective Actions
- Root Cause Analysis
- Corrections

Tools:

- Data Profiling Tools
- Business Rule Engine
- Data Parsing and Formatting
- Data Transformation and Standardization
- Data Enrichment
- Incident Management

Metrics:

- Return on Investment
- Levels of Quality
- Data Quality Scorecard
- Data Issues Reports
- Service Level Conformance
- Data Quality Plan Progress

(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 91 Context Diagram: Data Quality

1.1 Business Drivers

The business drivers for improving the trust in data through formal Data Quality Management are:

- Enhancing the stakeholder experience and organization's reputation
- Increasing the effectiveness of the organization
- Reducing risks and costs associated with poor quality data
- Improving organizational efficiency and productivity.

Data Quality Management achieves these outcomes using a systematic approach by:

- Understanding the context driving the need for the data (fitness)
- Matching the Data Quality needs to this context (purpose)
- Ensuring this quality is achieved repeatably through stable process.

The tangible benefits that organizations experience with data of sufficient quality include:

- Customers trusting the data to be correct and have confidence in dealing with the organization
- Employees answering questions more quickly and consistently
- Revenue gains through identifying business opportunities and effective invoicing
- Decreased customer service calls and ability to solve the calls when they arrive
- Staff spending less time trying to figure out if the data is right and more time using the data to gain insight, make good decisions and serve customers.
- The organization meeting complying with regulatory requirements and improving its credit rating.

Poor-quality data is risk-laden (see Chapter 1). It can damage an organization's reputation, resulting in fines, lost revenue, lost customers, and negative media exposure.

1.2 Goals and Principles

From a data user's perspective, "Every day I come to work and my data is ready. I do not have to spend hours (or days or months) correcting it, massaging it, or fixing business processes that have been damaged by low quality data. I understand the data's purposes and the data meets my needs." Data Quality Management focuses on these goals:

- Developing an approach to make data fit for purpose based on data consumers' requirements
- Defining standards and specifications for Data Quality controls as part of the data lifecycle
- Defining and implementing processes to measure, monitor, and report on Data Quality levels
- Identify and advocate for opportunities to improve the quality of data through process and system improvements.

Data Quality improvement programs should be guided by the following principles:

- **Criticality:** A Data Quality improvement program should focus on the data most critical to the enterprise and its customers. Priorities for improvement should be based on the criticality of the data and on the level of risk if data is not correct.

- **Standards-driven:** All stakeholders in the data lifecycle have Data Quality requirements. To the degree possible, these requirements should be defined in the form of measurable standards and expectations against which the quality of data can be measured.
- **Objective measurement and transparency:** Data Quality levels need to be measured objectively and consistently. Measurements and measurement methodology should be shared with stakeholders since they are the arbiters of quality.
- **Prevention:** The focus of Data Quality improvement program should be on preventing data errors and conditions that reduce the quality of data; it should not be focused on simply correcting records.
- **Root cause remediation:** Improving the quality of data goes beyond correcting errors. Problems with the quality of data should be understood and addressed at their root causes rather than just their symptoms. Because these causes are often related to process or system design, improving Data Quality often requires changes to processes and the systems that support them.
- **Embedded in business processes:** Business process owners are responsible for the quality of data produced through their processes. They must enforce Data Quality standards in their processes.
- **Systematically enforced:** System owners must systematically enforce Data Quality requirements.
- **Connected to service levels:** Data Quality reporting and issues management should be implemented and incorporated into Service Level Agreements (SLA).

1.3 Essential Concepts

1.3.1 Critical Data

Most organizations have a lot of data, not all of which is of equal importance. The first principle of Data Quality Management is to focus improvement efforts on data that is most important to the organization and its customers. Doing so gives the Data Quality improvement program scope and focus and enables it to make a direct, measurable impact on business needs.

Critical Data Elements are identified by linking them to the Data Quality business drivers; customer experience, effectiveness, and efficiency. Critical data needs to be managed to a defined Data Quality to meet these objectives. Without a particular data element there will be significant impairment to the success of the organization. This impairment size drives a business case for the level of Data Quality Management of a particular data element. This value prioritization approach creates a sliding scale of value. At a defined value level, data elements will become “critical data elements”.

For example, if the data in the customer email address field is incomplete, we will not be able to send product information to our customers via email and we will lose potential sales. We know that for every email we send out, we earn \$1.00 in revenue. This simple example links a clear value driver to Data Quality improvement.

The obvious question here might be, “can we have a scale of criticality?”. The answer is “yes”.

While specific drivers for criticality will differ by industry, there are common characteristics across organizations. Critical data is often used in:

- Regulatory, financial, or management reporting
- Business operational needs
- Measuring product quality and customer satisfaction
- Business strategy, especially efforts at competitive differentiation.

Master and Reference Data is usually critical by definition. Data sets or individual data elements can be assessed for criticality based on the processes that consume them, the nature of the reports they appear in, or the financial, regulatory, or reputational risk to the organization if something were to go wrong with the data.⁷²

1.3.2 Data Quality Dimensions

A Data Quality dimension is a measurable feature or characteristic of data. The term dimension is used to make the connection to dimensions in the measurement of physical objects (e.g., length, width, height). Data Quality dimensions provide a vocabulary for defining Data Quality requirements. From there, they can be used to define the results of the initial Data Quality Assessment as well as ongoing measurement. In order to measure the quality of data, an organization needs to establish dimensions that are both important to business processes (worth measuring) and measurable. Dimensions provide a basis for measurable rules, which themselves should be directly connected to potential risks in critical processes.

Data Quality dimensions are the key concepts that drive the Data Quality principles of “standards driven” and “objective measurement and transparency”. They are so useful, that when we ask, “What is high-Data Quality?”, we can answer with the Data Quality dimension requirements.

For example, we know that if the customer email address field is incomplete, then we will lose \$1.00 in revenue. After examining what is possible (and the reality of customers changing their email addresses), we determine that we can achieve 98% as a measure of Data Quality. We will measure and improve our processes until we have a valid, complete, and accurate email address for at least 98% of our customers.

While some practitioners use different sets of Data Quality dimensions (see section 8 Appendix of this chapter), the table below contains the common ideas. It contains definitions of a set of Data Quality dimensions, about which there is general agreement and describes approaches to measuring them.

⁷² See Jugulum (2014), Chapters 6 and 7 for an approach to rationalizing critical data.

Table 29 Common Dimensions of Data Quality

Dimension of Quality	Description
Validity	Validity refers to whether data values are consistent with a defined domain of values. A domain of values may consist of: • A data type (e.g., text with or without special characters, a number, or a date) • A data range (e.g., a numeric range, date range, or a set of valid text values in a reference table) • A format (e.g., a phone number must contain country, area codes, and the correct number of digits or a currency must have only two decimal places) • The precision of expected values. For example, a date-stamp must be recorded to a millisecond or a number recorded to no more than two decimal places. • The time frame (certain values may also only be valid for a specific length of time, for example, data that is generated from RFID (radio frequency ID) or some scientific data sets).
Completeness	Completeness refers to whether all the required data is present. For example, are all the mandatory components of an address populated, including the street number, street name, city, and country? For example, if the customer account is marked as a debt risk, there must be a debt manager appointed. If the customer account is not marked as a debt risk, there must not be a debt manager appointed. Completeness can be measured at the column, record, or data set level. • Are columns/data elements populated to the level expected? (Some columns are mandatory. Optional columns are populated only under specific conditions.) • Are records populated correctly? (Records with different statuses may have different expectations for completeness.) • Does the data set contain all the records expected?
Consistency	Consistency is ensuring that data values are coded using the same approach, assessment and valuation criteria. Consistency is between two different data items, potentially within a data set and between data sets, and across time. Consistency may be defined between one set of data element values and another data element set within the same record (record-level consistency), between one set of data element values and another data element set in different records (cross-record consistency), or between one set of data element values and the same data element set within the same record at different points in time (temporal consistency). Consistency can also be used to refer to consistency of format. Take care not to confuse consistency with accuracy or correctness. Consistency can be measured using the following ways: • Cross record consistency is between data values in the same column – are all the customer addresses recording the head-office details or do some record service delivery centers? • Between associated data sets, where data sets are expected to have values that are linked. For example, when moving data from data capture systems to data warehousing systems, a consistency quality measure will maintain confidence in the values in the data warehouse. • Across time can be applied to any of the previous cases. For example, has the grading of students been the same over time? If a subject is completed previously, would the same learning performance yield the same grade today?
Integrity	Integrity refers to the lack of incoherent values and broken relationships in data. Integrity can be measured using the following ways: • Coherence – where one data value implies a limited range in another data value and they are matching. For example, a country has a defined set of states or provinces. If a customer address is in a certain country, they must use a state or province name for the associated set. For example, Alberta is a province of Canada, Pisa is a province of Italy, and Tasmania is a state of Australia. • Parent-child (referential integrity) – where every child must have a parent data value. For example, all the customer address country names must exist on the valid list of countries where the organization is authorized to sell products to. Data sets without integrity are seen as corrupted or have data loss. Data sets without referential integrity have 'orphans' – invalid reference keys, or 'duplicates' – identical rows which may negatively affect aggregation functions.

Dimension of Quality	Description
Timeliness	Data Timeliness refers to the time that data needs to become accessible to a user after its capture or update. Timeliness is the expectation and actual delay before the data becomes available. For example, An electricity utility needs to maintain its power grid and must have demand data measured and available to a system operator within a few seconds. Otherwise, electricity generation may get out of alignment with demand, causing an energy excess or shortage. As another example, a statistical agency of government produces the Gross Domestic Product (GDP) report two months after the end of the quarter. In order to meet this timeliness requirement, all the data collections that make up the GDP will have their own specific timeliness requirements.
Currency	Currency is the date that the data was last updated relative to now and the likelihood that it is still correct. Different data sets will have different currency expectations from relatively static data to highly volatile data. Static data remains current for a long period. Volatile data remains current for a relatively short period. For example, country codes are relatively static, remaining current for a long period. A business' bank account is relatively volatile as the balance is constantly changing. Volatile data, like stock prices on financial web pages, will often be shown with an “as-of-time”, so that data consumers understand the risk that the data has changed since it was recorded. During the day, while the markets are open, such data will be updated frequently – it is volatile. Once markets close, the data will remain unchanged but will still be current since the market itself is inactive –it is static. Key terms used in currency include: • Update time – the time stamp of the last update • Volatility measures the rate of change of the date. It may measure across all the data values or measure within a segment. For example, the rate that domestic customers change their address may be different from the rate that corporate customers change their address. • Latency measures the time between when the data was created and when it was made available for use. For example, overnight population of a data warehouse may end up with different data latencies. Data can have a latency of 1 day for data entered into the system early on the prior day, but only a few minutes for data generated just before the load. (See Chapter 8.).
Reasonableness	Reasonableness asks whether a data pattern meets expectations. For example, whether a distribution of sales in a geographic area makes sense based on what is known about the customers in that area. For example, based on previous customer logins at 5pm, are today's customer logins to our systems out of the ordinary? Key terms used in reasonableness include: • Fixed values – this is where the measure is fixed, not dependent on previous behavior. The maximum currency value of a transaction may be limited in size by the technical limits of a bank's systems and other values bigger than this value is “unreasonable”. • Benchmark value – it is possible to measure the average, deviation or any other statistical function and ensure the latest set of values conforms to expected boundaries. For example, there have been between 5,000 and 10,000 customer address changes per day, but yesterday there were 2,000,000. On a government data collection, the average size company has 20 employees, but the latest returns had an unusual number with an average of 1,500 employees reported.
Uniqueness / Deduplication	Uniqueness states that no real-world entity exists more than once within the data set. Asserting uniqueness of the entities within a data set implies that each row relates to each unique entity in the real world, and only that specific entity. Uniqueness can be identified using the following techniques: • Key Structure – where there are duplicate keys within the data set (See Chapter 5). For example, in a data set containing customers, many of which share the same customer number. • Related data – where other data in the data may point to a duplicate. For example, two customers with different customer numbers, but having the same name, date of birth, and office address. • Inquiries or complaints that users cannot access certain data. For example, a customer with multiple services can only access some of their services as the others are recorded under a different customer number (which is a duplicate).

Dimension of Quality	Description
Accuracy	Accuracy refers to the degree that data correctly represents ‘real-life’ entities. For example, is the person’s name in our database actually the person’s name in real life? Does the customer actually use that email address? Accuracy is difficult to measure unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy use the other dimensions to imply accuracy. Common techniques to improve accuracy include: • Checking consistency with a data source that has been verified as accurate; a government company register (system of record) or a commercial company Information Broker (system of reference). • Ongoing calibration of devices with reality. For example, over time, weather gauges may drift from providing accurate measurements. An annual program of field inspections visits 10% of the devices and recalibrates them to a standard to ensure they are delivering accurate measurements. • Sampling data values with reality. For example, small number of emails are sent to customers (mainly) to drive sales, but also have the benefit of checking the accuracy of the email address (does it exist, is it used)?

1.3.3 Data Quality Business Rules

Business rules describe how organizations should operate internally in order to be successful and compliant with the outside world. Business Rules describe how data should exist in order to be useful and usable within an organization. Data Quality Rules are aligned with dimensions and are used to describe Data Quality requirements. For example, the business rules for managing state/province names could be:

- **Validity** – All state/province names must be a value from the reference table stored in the Metadata repository
- **Completeness** – All addresses must have a state/province if the country is divided into states/provinces. Otherwise, they should not have a state/province value
- **Integrity** – All state/province names must have a value that is linked to the country name of the address
- **Currency** – The reference values will be periodically updated and the current addresses will need to be brought up to date with the current names. This will keep all the state/province names current with the latest reference values.

Notes:

- Not every dimension is applicable every time
- This is very long handed, often business rules will be compressed to cover multiple Data Quality dimensions and multiple data elements at once
- Data Quality Rules can be implemented through a number of techniques, including data entry pick lists, reference data lookups, and quality inspection reports
- Data Quality Rules are commonly implemented in software or by using document templates for data entry.

Some common business rule types (aligned with most common dimensions) are:

- **Validity:** can be implemented through rules such as:

- **Format compliance rules:** One or more patterns specify values assigned to a data element, such as standards for formatting telephone numbers.
 - **Value domain membership:** Specify that a data element's assigned value is included in those enumerated in a defined data value domain, such as 2-character United States Postal Codes for a STATE field.
 - **Range conformance:** A data element assigned value must be within a defined numeric, lexicographic, or time range, such as greater than 0 and less than 100 for a numeric range.
 - **Mapping conformance:** Indicating that the value assigned to a data element must correspond to one selected from a value domain that maps to other equivalent corresponding value domain(s). The STATE data domain again provides a good example, since State values may be represented using different value domains (USPS Postal codes, FIPS 2-digit codes, full names), and these types of rules validate that 'AL' and '01' both map to 'Alabama.'
- **Completeness or Value presence and record completeness:** Rules defining the conditions under which missing values are acceptable or unacceptable.
- **Consistency rules:** Conditional assertions that refer to maintaining a relationship between two (or more) data elements based on the actual values of those data elements. For example, address validation where postal codes correspond to particular States or Provinces.
- **Consistency may also drive Definitional conformance rules:** Confirm that the same understanding of data definitions is implemented and used properly in processes across the organization. Confirmation includes algorithmic agreement on calculated fields, including any time or local constraints, and rollup and status interdependence rules.
- **Integrity rules:** Rules expressed as assertions to maintain a relationship between two (or more) data elements based on the actual values of those data elements. For example, address validation where postal codes correspond to particular States or Provinces.
- **Timeliness validation:** Rules that indicate the characteristics associated with expectations for accessibility and availability of data. Timeliness rules often break down a process into a series of stages where timeliness is monitored through the cycle to ensure the dependent data is also achieving the timeliness requirement.
- **Uniqueness:** Rules that specify which entities must have a unique representation and whether one and only one record exists for each represented real world object. In Data Quality checks, we may have rules that go beyond simple tests to assess the likelihood of a duplicate value. For example, it may search for people living at the same address that are behaving in a cooperative fashion (that leads us to believe they may be the same person).
- **Reasonableness:** May be implemented through aggregating functions applied to sets of data instances (see Section 4.5). Examples of aggregation checks include validating reasonableness by a:
 - *relative* measure. For example, today's record count in a file is assessed by keeping statistics over time to generate an understanding of "normality" which is used to assess reasonableness in this particular instance.

- *absolute* measure. For example, validating the reasonableness of an average amount calculated from a set of transactions by a set of fixed thresholds that the values must fall within.

These concepts of relative and absolute measures for determining reasonability can be used on any data types, such as an absolute date range (fixed period in time), or a relative date range (e.g. to today), computed totals, record counts, etc.

- **Accuracy verification:** Compare a data value against a corresponding value in a system of record or other verified source (e.g. marketing data purchased from a vendor) to verify that the values match.

1.3.4 Data Quality Improvement Lifecycle

Most approaches to improving Data Quality are based on the techniques of quality improvement in the manufacture of physical products.⁷³ In this paradigm data is understood as the product of a set of processes. At its simplest, a process is defined as a series of steps that turns inputs into outputs. A process that creates data may consist of one-step (data collection) or many steps: data collection, integration into a data warehouse, aggregation in a data mart, etc. At any step data can be negatively affected. It can be collected incorrectly, dropped or duplicated between systems, aligned or aggregated incorrectly, etc. Improving Data Quality requires the ability to assess the relationship between inputs and outputs in order to ensure that inputs meet the requirements of the process and that outputs conform to expectations. Since outputs from one process become inputs to other processes, requirements must be defined along the whole data chain.

A general approach to Data Quality improvement, shown in Figure 92, is a version of the Shewhart / Deming cycle.⁷⁴ Based on the scientific method, the Shewhart / Deming cycle is a problem-solving model known as ‘plan-do-check-act’. Improvement comes through a defined set of steps. The condition of the data must be measured against standards and, if it does not meet standards, root cause(s) of the discrepancy from standards must be identified and remediated. Root causes may be found in any of the steps of the process, technical or non-technical. Once remediated, data should be monitored to ensure that it continues to meet requirements.

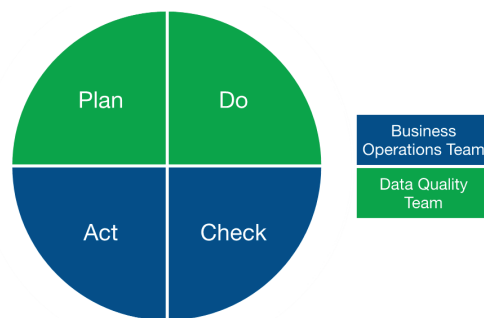


Figure 92 The Shewhart Chart with the role of a Data Quality Team indicated

⁷³ See Wang (1998), English (1999), Redman (2001), Loshin (2001), and McGilvray (2008). See Pierce (2004) for an overview of literature related to the concept of data as a product.

⁷⁴ See American Society for Quality: <http://bit.ly/1lelyBK> Plan-Do-Check-Act was originated by Walter Shewhart and popularized by W. Edwards Deming. 6 Sigma's Define, Measure, Analyze, Improve, Control (DMAIC) is a variation on this cycle.

For a given data set, a Data Quality Management cycle begins by the Data Quality team identifying the data that does not meet data consumers' requirements and data issues that are obstacles to the achievement of business objectives. Data needs to be assessed against key dimensions of quality and known business requirements. Root causes of issues will need to be identified so that stakeholders can understand the costs of remediation and the risks of not remediating the issues. This work is often done in conjunction with the relevant business operational teams, Data Stewards and other stakeholders.

In the *Plan* stage, the Data Quality team assesses the scope, impact, and priority of known issues and evaluates alternatives to address them. This plan should be based on a solid foundation of analysis of the root causes of issues. From knowledge of the causes and the impact of the issues, cost / benefit can be understood, priority can be determined, and a basic plan can be formulated to address them.

In the *Do* stage, the Data Quality team leads efforts to address the root causes of issues and plan for ongoing monitoring of data. For root causes that are based on non-technical processes, the Data Quality team can work with process owners to implement changes. For root causes that require technical changes, the Data Quality team should work with technical teams to ensure that requirements are implemented correctly and that technical changes do not introduce errors.

The *Check* stage involves actively monitoring the quality of data as measured against requirements. This is largely the responsibility of the business operations teams that manage the data on a daily basis. As long as data meets defined thresholds for quality, additional actions from the Data Quality team are not required. The processes will be considered under control and meeting business requirements. However, if the data falls below acceptable quality thresholds, then additional action must be taken to bring it up to acceptable levels.

The *Act* stage is for activities to address and resolve emerging Data Quality issues. This activity is usually also performed by the business operational team. The cycle restarts as the Data Quality team identifies new issues. Continuous improvement is achieved by starting a new cycle. New cycles begin as:

- Existing measurements fall below thresholds
- New data sets come under investigation
- New Data Quality requirements emerge for existing data sets
- Business rules, standards, or expectations change

The DMBOK Data Quality activities are mapped to this diagram in Table 30.

Table 30 Data Quality Activities

Dimension of Quality	Description
Plan (Data Quality Team)	1 Define High Quality Data 2 Define a Data Quality Policy 3 Define Scope of Initial Assessment 4 Perform Initial Data Quality Assessment 5 Identify & Prioritize Improvements 6 Define Goals for Data Quality Improvement
Do (Data Quality Team)	7 Develop and Deploy Data Quality Operations a. Develop Data Quality Operational Procedures

Dimension of Quality	Description
Check (Operations)	8. Develop and Deploy Data Quality Operations a. Correct Data Quality Defects b. Measure and Monitor Data Quality c. Report on Data Quality levels and findings
Act	Correct Data Quality Defects

This mapping enables understanding Data Quality using a manufacturing mindset.

1.3.5 Common Causes of Data Quality Issues

Many people assume that most Data Quality issues are caused by data entry errors and that adding more technology will solve the problem. A more sophisticated understanding recognizes that gaps in – or poor execution of – business and technical processes cause many more problems than mis-keying.

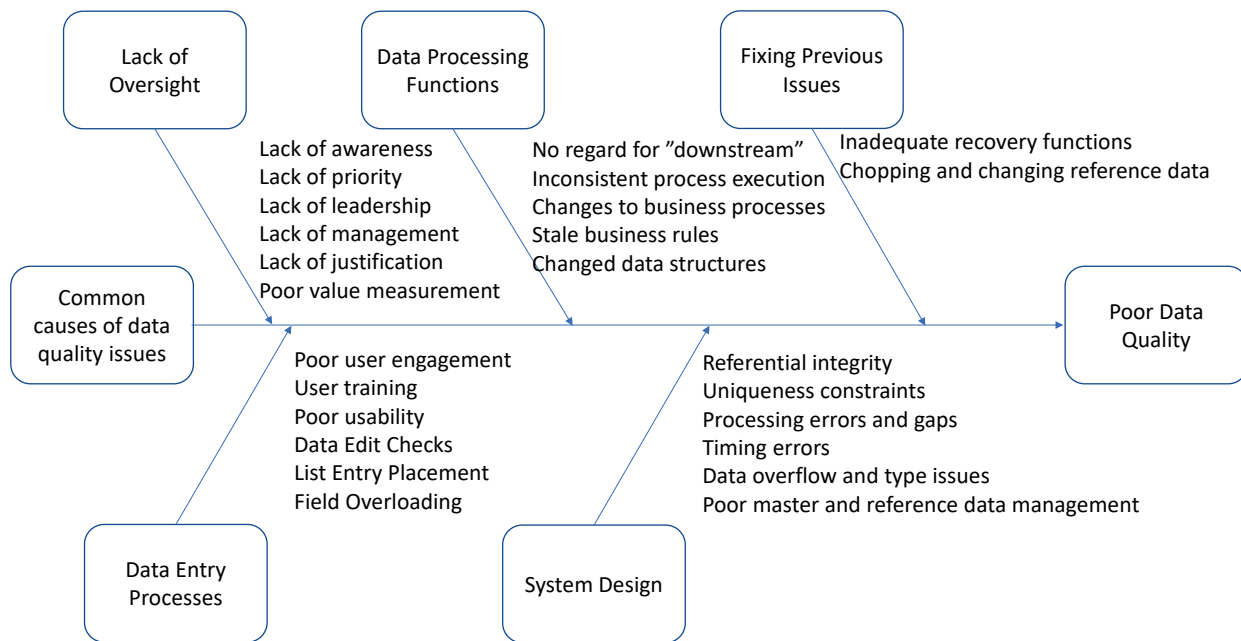


Figure 93 Sources of Data Quality Issues

Data Quality issues can emerge at any point in the data lifecycle – from creation to disposal. When investigating root causes, analysts should look for potential culprits like problems with data entry, data processing, system design and manual intervention in automated processes. Many issues will have multiple causes and contributing factors (especially if people have created ways to work around them). These causes of issues also imply ways to prevent issues: through improvement to interface design, testing of Data Quality Rules as part of processing, a focus on Data Quality within system design and strict controls on manual intervention in automated processes.

1.3.5.1 Issues Caused by Lack of Oversight

Many Data Quality problems are caused by a lack of organizational commitment to Data Quality in the form of both management and governance. For example, a business function may be solely managed to financial or completion metrics (such as finishing a call in 3 minutes or entering so many forms per day or just finishing the data entry in the next 2 weeks) without corresponding Data Quality metrics. This causes a lack of management focus and priority, as Data Quality is not valued equally with other performance measures. The existence of poor business metrics (e.g. missing data requirements) also demonstrates poor governance as there is no active oversight.

Every organization has information and data assets that are of value to its operations. Indeed the operations of every organization depend on the ability to share information. Despite this, few organizations manage these assets with rigor through the full lifecycle. Within most organizations data disparity (differences in data structure, format, and use of values) is a larger problem than just simple errors; it can be a major obstacle to the integration of data. One of the reasons data stewardship focus on defining terms and consolidating the language around data is because that is the starting point for improving the consistency of data.

Barriers to effective management of Data Quality include:⁷⁵

- Lack of awareness on the part of leadership and staff
- Lack of priority
- Lack of business governance
- Lack of leadership and management
- Difficulty in justification of improvements
- Inappropriate or ineffective instruments to measure value

These barriers have negative effects on customer experience, productivity, morale, organizational effectiveness, revenue, and competitive advantage. They increase the costs of running the organization and introduce risks as well. (See Chapter 16).

1.3.5.2 Issues Caused by Data Entry Processes

At the very start of the data lifecycle, data capture, there are a number of factors that can contribute to poor Data Quality. They include:

- **Poor user engagement:** If the people who are entering the data see little or no value in entering the data (even if they are paid), then Data Quality is likely to suffer. For example, if customers are asked to enter too much personal data to set up an account, they are likely to take less care or even deliberately respond inaccurately.
- **Training issues:** Lack of process knowledge can lead to incorrect data entry even if controls and edits are in place. If data producers are not aware of the impact of incorrect data or if they are incented for

⁷⁵ Adapted from *The Leader's Data Manifesto*. <https://dataleaders.org/>.

speed, rather than accuracy, they are likely to make choices based on drivers other than the quality of the data.

- **Poor usability:** User interfaces that are jumbled, inconsistent, lack clear instruction, use unexpected approaches, or ask for information in an unusual sequence require the user to work harder to enter the data. This is likely to lead to poorer Data Quality through both the inability to find the place to enter the data and omission through frustration.
- **No data entry edit checks:** If a data entry interface does not have edits or controls to prevent incorrect data from being put in the system data processors are likely to take shortcuts, such as skipping non-mandatory fields and failing to update defaulted fields. Key Data Quality dimensions managed in data entry are validity, completeness, consistency, integrity and reasonableness.
- **Inadequate list entry placement:** Even simple features of data entry interfaces, such as the order of values within a drop-down list, can contribute to data entry errors.
- **Field overloading:** Some organizations re-use fields over time for different business purposes rather than making changes to the data model and user interface. This practice results in inconsistent and confusing population of the fields.

1.3.5.3 Issues Caused by Data Processing Functions

As data progresses through a value chain, Data Quality issues may be introduced. The factors that cause this include:

- **No regard for “downstream” processes:** If each automated process and user works in a siloed environment without understanding where data comes from and where data goes to, then issues can arise with Data Quality. This may not even be deliberate. Users are often unaware of the use of their data and the issues they are causing.
- **Inconsistent business process execution:** Different user groups may use the data differently and perform business processes differently. Inconsistent execution may be due to training or documentation issues as well as to different requirements in different parts of the business.
- **Changes to business processes:** Business processes change over time. With these changes, new business rules and Data Quality requirements are introduced. Data is likely to be impacted unless changes to business rules are propagated throughout the entire system. Business rule changes are not always incorporated into systems in a timely or comprehensive manner.
- **Stale business rules:** Over time, business rules change. They should be periodically reviewed and updated. If there is automated measurement of rules then the technical process for measuring rules should also be updated. If it is not updated, issues may not be identified or false positives will be produced (or both).
- **Changed data structures:** Source systems may change structures without informing downstream consumers (both human and system) or without providing sufficient time to account for the changes.

This can result in invalid values or other conditions that prevent data movement and loading, or in more subtle changes that may not be detected immediately.

1.3.5.4 Issues Caused by System Design

Automated processing may have faults that introduce data errors that persist after they are fixed. There may also be choices made to encourage improved performance that reduce the Data Quality. Some of the System Design factors include:

- **Failure to enforce referential integrity:** Referential integrity is necessary to ensure high quality data at an application or system level. If referential integrity is not enforced or if validation is switched off (for example, to improve response times), various Data Quality issues can arise:
 - Duplicate data that breaks uniqueness rules
 - Orphan rows, which can be included in some reports and excluded from others, leading to multiple values for the same calculation
 - Inability to upgrade due to restored or changed referential integrity requirements
 - Inaccurate data due to missing data being assigned default values
- **Failure to enforce uniqueness constraints: Multiple copies of data instances within a table or file that is expected to contain unique instances** is an example of this failure. If there are insufficient checks for uniqueness of instances, or if the unique constraints are turned off in the database to improve performance, data aggregation results can be overstated.
- **Processing errors and gaps:** If the data mapping or layout is incorrect, or the rules for processing the data are not accurate, the data processed will have Data Quality issues, ranging from incorrect calculations to data being assigned to or linked to improper fields, keys, or relationships.
- **Timing errors:** If processing dependencies are not managed or processing continues despite a failure, systems could try to process data of different currency leading to data mismatch and data loss. For example, if a new electricity meter is installed on a house but the meter reading is not reset to the new meter's initial reading then electricity usage measurement errors will occur.
- **Data overflows and type issues:** If assumptions within the data model are not supported by the actual data then there will be Data Quality issues ranging from data loss due to field types and lengths being exceeded by the actual data, to data being assigned to improper IDs or keys.
- **Poor reference and Master Data Management:** By definition, reference and master data are shared between many business functions. Defects in data can have repercussions on many business operations. In particular, duplicates in the master list of clients are a common challenge. (see chapter 10).

1.3.5.5 Issues Caused by Fixing Issues

Manual data patches are changes made directly on the data in the database, not through the business rules in the application interfaces or processing. These are scripts or manual commands generally created in a hurry and used to 'fix' data in an emergency such as intentional injection of bad data, lapse in security, internal fraud, or external source for business disruption.

Like any untested code, they have a high risk of causing further errors through unintended consequences such as changing more data than required, or not propagating the patch to all historical data affected by the original issue. Most patches also change the data in place rather than preserving the prior state and adding corrected rows.

These changes are generally NOT undo-able without a complete restore from backup as there is only the database log to show the changes. Therefore, these shortcuts are strongly discouraged – they are opportunities for security breaches and business disruption longer than a proper correction would cause. All changes should go through a governed change management process.

Factors that drive the need for manual data patches on the database include:

- **Inadequate recovery functions:** If data becomes corrupted in the system there should be user functions able to recover the record and fix the errors. Without appropriate recovery functions the users are forced to fix the data using manual data patches.
- **Changing reference data:** Reference data should be changed infrequently. If reference data changes regularly data validity errors will arise in the database that may need to be corrected. If there are no functions to update reference data on the actual data then this will also drive manual data patches.

2. Activities

2.1 Define a Data Quality Framework

Improving Data Quality requires an approach that accounts for the work that needs to be done and the way people will execute it. Data Quality priorities must align with business strategy. Adopting or developing a framework and methodology will help guide both strategy and tactics while providing a means to measure progress and impacts. A framework should include methods to:

- Understand and prioritize business needs
- Identify the data critical to meeting business needs
- Define business rules and Data Quality standards based on business requirements
- Assess data against expectations
- Share findings and get feedback from stakeholders
- Prioritize and manage issues

- Identify and prioritize opportunities for improvement
- Measure, monitor, and report on Data Quality
- Manage Metadata produced through Data Quality processes
- Integrate Data Quality controls into business and technical processes

A framework should also account for how to organize for Data Quality and how to leverage Data Quality tools. As noted in the chapter introduction, improving Data Quality requires a Data Quality Function team to engage business and technical staff and define a program of work that addresses critical issues, defines best practices and puts in place operational processes that support ongoing management of Data Quality. Often such a team will be part of the Data Management Organization. Data Quality analysts will need to work closely with Data Stewards on all levels. They should also influence policy, including policy about business processes and systems development. However, such a team will not be able to solve all of an organization's Data Quality challenges. Data Quality work and a commitment to high quality data need to become embedded in organizational practices. The Data Quality Framework should account for how to extend best practices. (See Chapter 17.)

2.2 Define High Quality Data

Not all data is of equal importance. Data Quality Management efforts should focus first on the most important data in the organization: data that, if it were of higher quality, would provide greater value to the organization and its customers. Data can be prioritized based on factors such as regulatory requirements, financial value and direct impact on customers. Often, Data Quality improvement efforts start with Master Data, which is, by definition, among the most important data in any organization. The result of the importance analysis is a ranked list of data, which the Data Quality team can use to focus their work efforts.

Many people recognize poor quality data when they see it. Few are able to define what they mean by high quality data. Alternatively, they define it in very general terms: "The data has to be right." "We need accurate data." High quality data is fit for the purposes of data consumers. Before launching a Data Quality improvement program, it is beneficial to understand business needs, define terms, identify organizational pain points, and start to build consensus about the drivers and priorities for Data Quality improvement. Ask a set of questions to understand current state and assess organizational readiness for Data Quality improvement:

- What do stakeholders mean by 'high quality data'?
- What is the impact of low quality data on business operations and strategy?
- How will higher quality data enable business strategy?
- What priorities drive the need for Data Quality improvement?
- What is the tolerance for poor quality data?
- What governance is in place to support Data Quality improvement?
- What additional governance structures will be needed?

Getting a comprehensive picture of the current state of Data Quality in an organization requires approaching the question from different perspectives:

- An understanding of business strategy and goals
- Interviews with stakeholders to identify pain points, risks, and business drivers

- Direct assessment of data through profiling and other form of analysis
- Documentation of data dependencies in business processes
- Documentation of technical architecture and systems support for business processes

This kind of assessment can reveal a significant number of opportunities. These need to be prioritized based on the potential benefit to the organization. Using input from stakeholders, including Data Stewards and business and technical Subject matter experts, the Data Quality team should define the meaning of Data Quality and propose priorities.

2.3 Identify Dimensions and Supporting Business Rules

Having identified the critical data, Data Quality analysts need to identify business rules that describe or imply expectations about the dimensions of Data Quality. Often rules themselves are not explicitly documented. They may need to be reverse engineered through analysis of existing business processes, workflows, regulations, policies, standards, system edits, software code, triggers and procedures, status code assignment and use, and plain old common sense. For example, if a marketing company wants to target efforts at people in a specific demographic then potential indexes of Data Quality might be the level and reasonability of population in demographic fields like birth date, age, gender and household income.

Most business rules are associated with how data is collected or created, but Data Quality measurement centers around whether data is fit for use. The two (data creation and data use) are related. People want to use data because of what it represents and why it was created. For example, to understand an organization's sales performance during a specific quarter or over time depends on having reliable data about the sales process (number and type of units sold, volume sold to existing customers vs. new customers, etc.).

It is not possible to know all the ways that data might be used, but it is possible to understand the process and rules by which data was created or collected. Measurements that describe whether data is fit for use should be developed in relation to known uses and measurable rules based on dimensions of Data Quality: completeness, validity, integrity, etc. that provide the basis for meaningful metrics. Dimensions of quality enable analysts to characterize both rules (field X is mandatory and must be populated) and findings (e.g., the field is not populated in 3% of the records; the data is only 97% complete).

At the field or column level, rules can be straightforward. Completeness rules are a reflection of whether a field is mandatory or optional, and, if optional, the conditions under which it should be populated. Validity rules are dependent on stipulating the domain of valid values and, in some cases, the relationship between fields. For example, a US ZIP Code needs to be valid, in and of itself, and correctly associated with a US State code. Rules should also be defined at the data set level. For example, every customer must have a valid mailing address.

Defining Data Quality Rules is challenging because most people are not used to thinking about data in terms of rules. It may be necessary to get at the rules indirectly by asking stakeholders about the input and output requirements of a business process. It also helps to ask about pain points, what happens when data is missing or incorrect, how they identify issues, how they recognize bad data, etc. Keep in mind that it is not necessary to know all the rules in order to assess data. Discovery and refinement of rules is an ongoing process. One of the best ways to get at rules is to share results of assessments. These results often give stakeholders a new

perspective on the data from which they can articulate rules that tell them what they need to know about the data.

2.4 Perform an Initial Data Quality Assessment

Once the most critical business needs and the data that supports them have been identified, the most important part of the Data Quality Assessment is actually looking at that data, querying it to understand data content and relationships, and comparing actual data to rules and expectations. The first time this is done, analysts will discover many things: undocumented relationships and dependencies within the data, implied rules, redundant data, contradictory data, etc., as well as data that actually does conform to rules. With the help of data stewards, other subject matter experts and data consumers, Data Quality analysts will need to sort out and prioritize findings.

- The goal of an initial Data Quality Assessment is to learn about the data in order to define an actionable plan for improvement. It is usually best to start with a small, focused effort – a basic proof of concept – to demonstrate how the improvement process works. Steps include:
 - Define the goals of the assessment; these will drive the work
 - Identify the data to be assessed; focus should be on a small data set, even a single data element, or a specific Data Quality problem
 - Identify uses of the data and the consumers of the data
 - Identify known risks with the data to be assessed, including the potential impact of data issues on organizational processes
 - Inspect the data based on known and proposed rules
 - Document levels of non-conformance and types of issues
 - Perform additional, in-depth analysis based on initial findings in order to
 - Quantify findings
 - Prioritize issues based on business impact
 - Develop hypotheses about root causes of data issues
 - Meet with data stewards, subject matter experts, and data consumers to confirm issues and priorities
 - Use findings as a foundation for planning:
 - Remediation of issues, ideally at their root causes
 - Controls and process improvements to prevent issues from recurring
 - Ongoing controls and reporting

2.5 Identify and Prioritize Potential Improvements

Having demonstrated that Data Quality should be improved, the next goal is to do it strategically. Doing so requires identifying and prioritizing potential improvements. Identification may be accomplished by full-scale data profiling of larger data sets to understand the breadth of existing issues. It may also be accomplished by other means such as interviewing stakeholders about the data issues that impact them and following up with analysis of the business impact of those issues. Ultimately, prioritization requires a combination of data analysis and discussion with stakeholders.

The steps to perform a full data profiling and analysis are essentially the same as those in performing a small-scale assessment: define goals, understand data uses and risks, measure against rules, document and confirm findings with subject matter experts, use this information to prioritize remediation and improvement efforts. However, there are sometimes technical obstacles to full-scale profiling. The effort will need to be coordinated across a team of analysts and overall results will need to be summarized and understood if an effective action plan is to be put in place. Large-scale profiling efforts, like those on a smaller scale, should still focus on the most critical data.

2.6 Define Goals for Data Quality Improvement

The knowledge obtained through the preliminary assessments forms the basis for specific Data Quality improvement program goals. Improvement can take different forms, from simple remediation (e.g., correction of errors on records) to remediation of root causes. Remediation and improvement plans should account for quick hits – issues that can be addressed immediately at low cost – and longer-term strategic changes. The strategic focus of such plans should be to address root causes of issues and to put in place mechanisms to prevent issues in the first place.

Be aware that many things can get in the way of improvement efforts: system constraints, age of data, ongoing project work that uses the questionable data, overall complexity of the data landscape, cultural resistance to change. To prevent these constraints from stalling the program, set specific, achievable goals based on consistent quantification of the business value of the improvements to Data Quality.

For example, a goal may be to improve the completeness of customer data from 90% to 95% based on process improvements and system edits. Obviously, showing improvement will involve comparing initial measurements and improved results. But the value comes with benefits of the improvement: fewer customer complaints, less time spent correcting errors, etc. Measure these things to explain the value of the improvement work. No one cares about levels of field completeness unless there is a business impact. There must be a positive return on investment for improvements to data. When issues are found, determine ROI of fixes based on:

- The criticality (importance ranking) of the data affected
- Amount of data affected
- The age of the data
- Number and type of business processes impacted by the issue
- Number of customers, clients, vendors, or employees impacted by the issue
- Risks associated with the issue
- Costs of remediating root causes
- Costs of potential workarounds

In assessing issues, especially those where root causes are identified and technical changes are required, always seek out opportunities to prevent issues from recurring. Preventing issues generally costs less than correcting them – sometimes orders of magnitude less.

2.7 Develop and Deploy Data Quality Operations

Many Data Quality improvement programs get started through a set of improvement projects identified via results of the Data Quality Assessment. In order to sustain Data Quality, a Data Quality Function should put in place a process that allows the team to manage Data Quality Rules and standards, monitor data's ongoing conformance with rules, identify and manage Data Quality issues and report on quality levels. In support of these activities, Data Quality analysts and data stewards will also be engaged in activities such as documenting data standards and business rules and establishing Data Quality requirements for vendors.

2.7.1 Manage Data Quality Rules

The process of profiling and analyzing data will help an organization discover (or reverse engineer) business and Data Quality Rules. As the Data Quality practice matures, the capture of such rules should be built into the system development and enhancement process. Defining rules upfront will:

- Set clear expectations for Data Quality dimensions
- Provide requirements for system edits and controls that prevent data issues from being introduced
- Provide Data Quality requirements for all relevant dimensions to vendors and other external parties
- Create the foundation for ongoing Data Quality measurement and reporting

In short, Data Quality Rules and standards are a critical form of Metadata. To be effective, they need to be managed as Metadata. Rules should be:

- **Documented consistently:** Establish standards and templates for documenting rules so that they have a consistent format and meaning.
- **Defined in terms of Data Quality dimensions:** Dimensions of quality help people understand what is being measured. Consistent application of dimensions will help with the measurement and issue management processes.
- **Tied to business impact:** While Data Quality dimensions enable understanding of common problems, they are not a goal in-and-of-themselves. Standards and rules should be connected directly to their impact on organizational success. Measurements that are not tied to business processes should not be taken.
- **Backed by data analysis:** Data Quality analysts should not guess at rules. Rules should be tested against actual data. In many cases, rules will show that there are issues with the data. But analysis can also show that the rules themselves are not complete.
- **Confirmed by subject matter expert:** The goal of the rules is to describe how the data should look. Often, it takes knowledge of organizational processes to confirm that rules correctly describe the data. This knowledge comes when subject matter experts confirm or explain the results of data analysis.
- **Accessible to all data consumers:** All data consumers should have access to documented Data Quality Rules. Such access allows them to better understand the data. It also helps to ensure that the

rules are correct and complete. Ensure that consumers have a means to ask questions about and provide feedback on rules.

2.7.2 Measure and Monitor Data Quality

- Much of the work of a Data Quality team will focus on measuring and reporting on quality. High-level categories of Data Quality metrics include:
 - Return on Investment: Statements on cost of improvement efforts vs. the benefits of improved Data Quality
 - Levels of quality: Measurements of the number and percentage of errors or requirement violations within a data set or across data sets
 - Data Quality trends: Quality improvement over time (i.e., a trend) against thresholds and targets, or quality incidents per period
 - Data issue management metrics:
 - Counts of issues by dimensions of Data Quality
 - Issues per business function and their statuses (resolved, outstanding, escalated)
 - Issue by priority and severity
 - Time to resolve issues
 - Data Quality plan rollout: status of the rollout, next stages and an ongoing prioritization process to maintain value in the roadmap.

The operational Data Quality Management procedures depend on the ability to measure and monitor the quality of data. There are two equally important reasons to implement operational Data Quality measurements:

1. To inform data consumers about levels of quality
2. To manage risk that change may be introduced through changes to business or technical processes

Some measurements serve both purposes. Measurements should be developed based on findings from data assessment and root cause analysis. Measurements intended to inform data consumers will focus on critical data elements and relationships that, if they are not of high quality, will directly impact business processes. Measurements related to managing risk should focus on relationships that have gone wrong in the past and may go wrong in the future. For example, if data is derived based on a set of ETL (extract, transform, load) rules and those rules may be impacted by changes to business processes, measurements should be put in place to detect changes to the data.

Knowledge of past problems should be applied to manage risk. For example, if numerous data issues are associated with complex derivations, then all derivations should be assessed – even those that have not been associated with data issues. In most cases, it is worthwhile to put in place measurements that monitor functions similar to those that have had problems.

In Figure 94, the Data Quality threshold is the dotted line in each graph. The Data Quality team progresses each of these trends:

- 2, 4, and 5 are ongoing issues (the team is monitoring progress to bring them back under threshold)
- 1, 3, 10 are today's issues (the team is mobilizing efforts to identify and address the trend)
- 7, 8, 11 are on today's watch list (the team is taking an interest but has not yet mobilized to address)
- 6 is a long-term intransigent issue (the team has tried a couple of ideas but is still looking for a permanent solution)
- 9, 12 are fine (the team are happy with these trends).

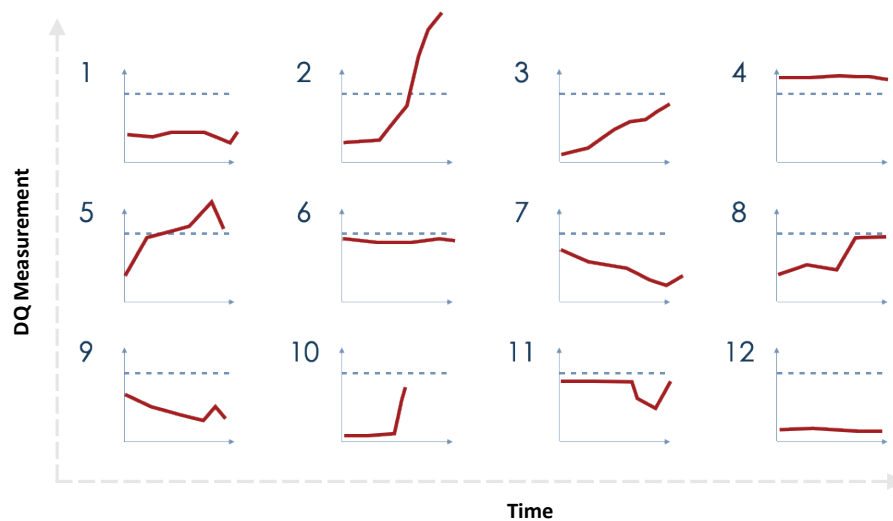


Figure 94 Illustrations of data defects trends over time

Table 31 Data Quality Monitoring Techniques

Granularity	Common Dimensions	Treatment
Data Element	Validity, Integrity, Consistency, Reasonableness	Edit checks in application Data element validation services Specially programmed applications
Data Record	Consistency, completeness, currency	Edit checks in application Data record validation services Specially programmed applications
Data set	Completeness, Uniqueness, Reasonableness	Inspection inserted between processing stages

2.7.3 Develop Procedures for Managing Data Issues

In modern enterprises, many different types of incidents are already effectively managed. For example, health and safety issues, property and building issues, and Information Technology issues. Therefore, before developing new processes for managing data issues, look to extend an existing process and toolset. The key elements of an incident management process are:

- **Standardize Data Quality issues and activities:** Since the terms used to describe data issues may vary across lines of business, it is valuable to define a standard vocabulary for the concepts used. Doing so will simplify classification and reporting. Standardization also makes it easier to measure the volume of issues and activities, identify patterns and interdependencies between systems and participants, and report on the overall impact of Data Quality activities. The classification of an issue may change as the investigation deepens and root causes are exposed.
- **Provide an assignment process for data issues:** The operational procedures direct the analysts to assign Data Quality incidents to individuals for diagnosis and to provide alternatives for resolution. Drive the assignment process within the incident tracking system by suggesting those individuals with specific areas of expertise.
- **Manage issue escalation procedures:** Data Quality issue handling requires a well-defined system of escalation based on the impact, duration, or urgency of an issue. Specify the sequence of escalation within the Data Quality Service Level Agreement. The incident tracking system will implement the escalation procedures, which helps expedite efficient handling and resolution of data issues.
- **Manage Data Quality resolution workflow:** The Data Quality SLA specifies objectives for monitoring, control, and resolution - all of which define a collection of operational workflows. The incident tracking system can support workflow management to track progress with issues diagnosis and resolution.

Even when using a mature, standardized process, the Data Quality team must design and implement detailed operational procedures for the steps within this:

- **Diagnosing issues:** The objective is to review the symptoms of the Data Quality incident, trace the lineage of the data in question, identify the problem and where it originated and pinpoint potential root causes of the problem. The procedure should describe how the Data Quality Operations team would:
 - Review the data issues in the context of the appropriate information processing flows and isolate the location in the process where the flaw is introduced
 - Evaluate whether there have been any environmental changes that would cause errors entering into the system
 - Evaluate whether or not there are any other process issues that contributed to the Data Quality incident
 - Determine whether there are issues with external data that have affected the quality of the data

NOTE: The work of root cause analysis requires input from technical and business subject matter experts. While the Data Quality team may lead and facilitate this kind of work effort, success requires cross-functional collaboration.

- **Formulating options for remediation:** Based on the diagnosis, evaluate alternatives for addressing the issue. These may include:

- Addressing non-technical root causes such as lack of training, lack of leadership support, unclear accountability and ownership, etc.
 - Modification of the systems to eliminate technical root causes
 - Developing controls to prevent the issue
 - Introducing additional inspection and monitoring
 - Directly correcting flawed data
 - Taking no action based on the cost and impact of correction versus the value of the data correction
- **Resolving issues:** Having identified options for resolving the issue the Data Quality team must confer with the business data owners to determine the best way to resolve the issue. These procedures should detail how the analysts:
 - Assess the relative costs and merits of the alternatives
 - Recommend one of the planned alternatives
 - Provide a plan for developing and implementing the resolution
 - Implement the resolution

2.7.4 Establish Data Quality Service Level Agreements

A Data Quality Service Level Agreement (SLA) specifies an organization's expectations for response and remediation for Data Quality issues at key boundaries. Key boundaries include:

- Entry and exit from the organization: For example, a supplier is providing critical data to an organization for use in their data processing, so a key part of the contract is the Data Quality requirements.
- Key separation points within an organization: For example, a research unit within an organization will acquire data from various locations, both internal and external to the organization.

Data Quality inspections as scheduled in the SLA help to identify issues to fix, and over time, reduce the number of issues. While enabling the isolation and root cause analysis of data flaws, there is an expectation that the operational procedures will provide a scheme for remediation of root causes within an agreed timeframe. Having Data Quality inspection and monitoring in place increases the likelihood of detection and remediation of a Data Quality issue before a significant business impact can occur. Data Quality control defined SLA includes the same information as Data Quality requirements (see section ???) with the following additions:

- Data elements covered by the agreement
- Business impacts associated with data flaws
- Data Quality dimensions and business rules associated with each data element
- Methods for measuring against those expectations
- Acceptability threshold for each measurement
- Steward(s) to be notified in case the acceptability threshold is not met
- Timelines and deadlines for expected resolution or remediation of the issue
- Escalation strategy, and possible rewards and penalties

The Data Quality SLA also defines the roles and responsibilities associated with performance of operational Data Quality procedures. Data stewards and the operational Data Quality staff, while upholding the level of Data Quality service, should consider their Data Quality SLA constraints and connect Data Quality to individual performance plans.

When issues are not addressed within the specified resolution times an escalation process must exist to communicate non-observance of the level of service up the management and governance chain. The Data Quality SLA establishes the time limits for notification generation, the names of those in that management chain and when escalation needs to occur. Given the set of Data Quality Rules, methods for measuring conformance, the acceptability thresholds defined by the business clients, and the service level agreements, the Data Quality team can monitor compliance of the data to the business expectations, as well as how well the Data Quality team performs on the procedures associated with data errors. SLA reporting can be on a scheduled basis driven by business and operational requirements. Particular focus will be on report trend analysis in cases focused on periodic rewards and penalties if such concepts are built into the SLA framework.

2.7.5 Data Quality Response

The work of assessing the quality of data and managing data issues will not benefit the organization unless the information is shared through reporting so that data consumers understand the condition of the data. Reporting should focus around:

- Data Quality scorecard which provides a high-level view of the scores associated with various metrics, reported to different levels of the organization within established thresholds
- Data Quality trends which show over time how the quality of data is measured and whether trending is up or down
- SLA Metrics such as whether operational Data Quality staff diagnose and respond to Data Quality incidents in a timely manner
- Data Quality issue management which monitors the status of issues and resolutions
- Conformance of the Data Quality team to governance policies
- Conformance of IT and business teams to Data Quality policies
- Positive effects of improvement projects

Reporting should align to metrics in the Data Quality SLA as much as possible so that the team's goals are aligned with those of its customers. The Data Quality improvement program should also report on the positive effects of improvement projects. It is best to do this in business terms to continually remind the organization of the direct effect that data has on customers.

3. Tools

While the focus of Data Quality improvement efforts is often on the prevention of errors, Data Quality can also be improved through some forms of data processing. (See Chapter 8.) Tools should be selected and tool architectures should be set in the planning phase of the enterprise Data Quality improvement program. Tools

provide a partial rule set starter kit but organizations need to create and input their own context specific rules and actions into any tools. The key functions of tools are:

- Profiling: exploring patterns of data values
- Business Rule Engines: specialist tools that automate the handling of Data Quality requirements
- Data Parsing and formatting: breaking data values into components and reassembling them to identify a correct response
- Data Transformation and Standardization: Used to change the values to meet standard requirements
- Data Enrichment: Used to augment the data with additional data (maybe thought of Metadata)
- Incident Management: Used to automate data issue management.

3.1 Data Profiling Tools

Data profiling tools produce high-level statistics that enable analysts to identify patterns in data and perform initial assessment of quality characteristics. Some tools can be used to perform ongoing monitoring of data. Profiling tools are particularly important for data discovery efforts because they enable assessment of large data sets. Profiling tools augmented with data visualization capabilities will aid in the process of discovery. (See Chapters 5 and 8, and Section 1.3.9.).

Data profiling is only the first step in data analysis. It helps identify potential issues. Data Quality team members also need to query data more deeply to answer questions raised by profiling results and find patterns that provide insight into root causes of data issues. For example, querying to discover and quantify other aspects of Data Quality, such as uniqueness and integrity.

3.2 Business Rule Templates and Engines

Business Rule Templates are specialist tools that allow analysts to capture expectations for data standard forms. Templates help bridge the communications gap between business and technical teams. Consistent formulation of rules makes it easier to translate business needs into code, whether that code is embedded in a rules engine, the data analyzer component of a data-profiling tool, or a data integration tool. A template can have several sections, one for each type of business rule to implement.

3.3 Data Parsing and Formatting

Data Parsing is the process of analyzing data using pre-determined rules to define its content or value. Data parsing enables the data analyst to define sets of patterns that feed into a rule engine used to distinguish between valid and invalid data values. Matching specific pattern(s) triggers actions.

Many Data Quality issues involve situations where variation in data values representing similar concepts introduces ambiguity. The separate components (commonly referred to as ‘tokens’) can be extracted and rearranged into a standard representation to create a valid pattern. When an invalid pattern is recognized, the

application may attempt to transform the invalid value into one that meets the rules. Perform standardization by mapping data from some source pattern into a corresponding target representation.

Another good example is a customer name since names may be represented in thousands of different forms. A good standardization tool will be able to parse the different components of a customer name, such as given name, middle name, family name, initials, titles, generational designations and then rearrange those components into a standard representation that other data services will be able to manipulate.

The human ability to recognize familiar patterns contributes to an ability to characterize variant data values belonging to the same abstract class of values; people recognize different types of telephone numbers because they conform to frequently used patterns. An analyst describes the format patterns that all represent a data object, such as Person Name, Product Description, and so on. A Data Quality tool parses data values that conform to any of those patterns, and even transforms them into a single, standardized form that will simplify the assessment, similarity analysis, and remediation processes. Pattern-based parsing can automate the recognition and subsequent standardization of meaningful value components.

3.4 Data Transformation and Standardization

During normal processing, data rules trigger and transform the data into a format that is readable by the target architecture. However, readable does not always mean acceptable. Rules are created directly within a data integration stream or rely on alternate technologies embedded in or accessible from within a tool. Data transformation builds on these types of standardization techniques. Organizations should guide rule-based transformations by mapping data values in their original formats and patterns into a target representation. Parsed components of a pattern are subjected to rearrangement, corrections, or any changes as directed by the rules in the knowledge base. In fact, standardization is a special case of transformation, employing rules that capture context, linguistics and idioms recognized as common over time, through repeated analysis by the rules analyst or tool vendor. (See Chapter 3.)

3.5 Data Enrichment

Data enrichment or enhancement is the process of adding data elements to a data set to increase its quality, including usability. Some enhancements are gained by integrating data sets internal to an organization. External data can also be purchased to enhance organizational data (see Chapter 10). Examples of data enhancement include:

- **Time/Date stamps:** One way to improve data is to document the time and date that data items are created, modified, or retired, which can help to track historical data events. If issues are detected with the data, timestamps can be very valuable in root cause analysis because they enable analysts to isolate the timeframe of the issue.
- **Audit data:** Auditing can document data lineage which is important for historical tracking as well as validation.
- **Reference vocabularies:** Business specific terminology, ontologies and glossaries enhance understanding and control while bringing customized business context.

- **Contextual information:** Adding context such as location, environment, or access methods and tagging data for review and analysis.
- **Geographic information:** Geographic information can be enhanced through address standardization and geocoding, which includes regional coding, municipality, neighborhood mapping, latitude / longitude pairs, or other kinds of location-based data.
- **Demographic information:** Customer data can be enhanced through demographic information such as age, marital status, gender, or income. Business entity data can be associated with annual revenue, number of employees, size of occupied space etc.
- **Psychographic information:** Data used to segment the target populations by specific behaviors, habits, or preferences, such as product and brand preferences, organization memberships, leisure activities, commuting transportation style, shopping time preferences etc.
- **Valuation information:** Use this kind of enhancement for asset valuation, inventory, and sale.

3.6 Incident Management System

Decisions made during the issue management process should be tracked in an Incident Management System. When the data in such a system is managed well it can provide valuable insight into the causes and costs of data issues. Include a description of the issue and the root causes, options for remediation and the decision on how to resolve the issue.

The Incident Management System will collect performance data relating to issue resolution, work assignments, volume of issues, frequency of occurrence, as well as the time to respond, diagnose, plan a solution, and resolve issues. These metrics can provide valuable insights into the effectiveness of the current workflow, as well as systems and resource utilization. They are important management data points that can drive continuous operational improvement for Data Quality control.

Incident management data also helps data consumers. Decisions based upon remediated data should be made with knowledge that it has been changed, why it has been changed and how it has been changed. That is one reason why it is important to record the methods of modification and the rationale for them. Make this documentation available to data consumers and developers researching code changes. While changes may be obvious to the people who implement them, the history of changes will be lost to future data consumers unless it is documented.

4. Techniques

Data Quality techniques are arranged into a standard sequence:

- Metrics –effectively measuring the quality of data
- Profiling – finding statistical and relationship patterns
- Preventative or corrective actions – how to avoid Data Quality problems in the first place
- Root cause analysis – getting to the fundamentals of a problem
- Corrections – how to tidy up a data issue

4.1 Effective Metrics

A critical component of managing Data Quality is developing metrics that inform data consumers about quality characteristics that are important to their uses of data. Many things can be measured, but not all of them are worth the time and effort. In developing metrics, Data Quality analysts should account for these characteristics:

- **Measurability:** A Data Quality metric needs to be something that can be counted. For example, address completeness needs to be defined in order to be measured. Expected results should be quantifiable within a discrete range.
- **Business relevance:** While many things are measurable, not all translate into useful metrics. Measurements need to be relevant to data consumers. The value of the metric is limited if it cannot be related to some aspect of business operations or performance. Every Data Quality metric should correlate with the influence of the data on key business expectations.
- **Acceptability:** The Data Quality dimensions frame the business requirements for Data Quality. Quantifying along the identified dimension provides hard evidence of Data Quality levels. Determine whether data meets business expectations based on specified acceptability thresholds. If the score is equal to or exceeds the threshold, the quality of the data meets business expectations. If the score is below the threshold, it does not.
- **Accountability / Stewardship:** Metrics should be understood and approved by key stakeholders (e.g., business owners and Data Stewards). They are notified when the measurement for the metric shows that the quality does not meet expectations. The business data owner is accountable, while a data steward takes appropriate corrective action.
- **Controllability:** A metric should reflect a controllable aspect of the business. In other words, if the metric is out of range, it should trigger action to improve the data. If there is no way to respond, then the metric is probably not useful.
- **Trending:** Metrics enable an organization to measure Data Quality improvement over time. Tracking helps Data Quality team members monitor activities within the scope of a Data Quality SLA and data sharing agreement and demonstrate the effectiveness of improvement activities. Once an information process is stable, statistical process control techniques can be applied to detect changes to the predictability of the measurement results and the business and technical processes on which it provides insight.

Metric measurement results can be described at two levels: the detail related to the execution of individual rules and overall results aggregated from the rules. Each rule should have a standard, target, or threshold index for comparison. This function most often reflects the percentage of correct data or percentage of exceptions depending on the formula used. For example:

$$ValidDQI(r) = \frac{(TextExecutions(r) - ExceptionsFound(r))}{TestExections(r)}$$

$$InvalidDQI(r) = \frac{ExceptionsFound(r)}{TestExections(r)}$$

r represents the rule being tested. For example, 10,000 tests of a business rule (r) found 560 exceptions. In this example, the Valid Data Quality result would be $9440/10,000 = 94.4\%$, and the Invalid Data Quality result would be $560/10,000 = 5.6\%$.

Organizing the metrics and results as shown in Table 30 can help to structure measures, metrics and indicators across the report, reveal possible rollups and enhance communications. The report can be more formalized and linked to projects that will remediate the issues. Filtered reports are useful for data stewards looking for trends and contributions. Table 30 provides examples of rules constructed in this manner. Where applicable, results of rules are expressed in both positive percentages (the portion of the data that conforms to rules and expectations) and negative percentages (the portion of the data that does not conform to the rule).

Data Quality Rules provide the foundation for operational management of Data Quality. Rules can be integrated into application services or data services that supplement the data lifecycle, either through Commercial Off The Shelf (COTS) Data Quality tools, rules engines and reporting tools for monitoring and reporting, or custom-developed applications.

Table 32 Data Quality Metric Examples

Example of completeness

Business Rule	General	Example
Business Rule	Population of field is mandatory	Postal Code must be populated in the address table for 100%.
Measure	Count the number of records where data is populated, compare to the total number of records	Count populated: 700,000 Count not populated: 300,000 Total count: 1,000,000
Metrics (calculation)	Divide the obtained number of records where data is populated by the total number of records in the table or database and multiply it by 100 to get to percentage complete	Positive measure: $700,000/1,000,000 \times 100 = 70\%$ populated Negative measure: $300,000/1,000,000 \times 100 = 30\%$ not populated
Status indicator	Acceptable if the business rule is met. Unacceptable is the business rule is not met.	Unacceptable. There are too much unpopulated Postal Codes.

Example of uniqueness

Business Rule	General	Example
Business Rule	There should be only one record per entity instance in a table	There should be one and only one current row per postal code on the Postal Codes master list: 0% duplicates.
Measure	Count the number of duplicate records identified; report on the percentage of records that represent duplicates	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000
Metrics (calculation)	Divide the number of duplicate records by the total number of records in the table or database and multiply it by 100	$10,000/1,000,000 \times 100 = 1.0\%$ of postal codes are present on more than one current row
Status indicator	Acceptable if the business rule is met. Unacceptable is the business rule is not met.	Unacceptable. There are too many duplicates.

Example of timeliness

Business Rule	General	Example
Business Rule	Records must arrive within a scheduled timeframe	99% of equity market record should arrive within 5 minutes of being transacted
Measure	Count the number of records failing to arrive on time from a data service for business transactions to be completed	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000

Business Rule	General	Example
Metrics (calculation)	Divide the number of incomplete transactions by the total number of attempted transactions in a time period and multiply by 100.	Positive: $(1,000,000 - 2000) / 1,000,000 * 100 = 99.8\%$ of transaction records arrived within defined timeframe Negative: $2000 / 1,000,000 * 100 = 0.20\%$ of transactions did not arrive within defined timeframe
Status indicator	Acceptable if the business rule is met. Unacceptable if the business rule is not met.	Acceptable. More than 99% of records arrive within 5 minutes.

Example of validity

Business Rule	General	Example
Business Rule	If field X = value 1, then field Y must = value 1-prime	Alle shipped orders should be billed
Measure	Count the number of records where the rule is met	Count of records where status for shipping = 'Shipped' and status for billing = 'Billed': 999,000 Count of total records: 1,000,000
Metrics (calculation)	Divide the number of records that meet the condition by the total number of records	Positive: $999,000 / 1,000,000 * 100 = 99.9\%$ of records conform to the rule Negative: $(1,000,000 - 999,000) / 1,000,000 * 100 = 0.1\%$ do not meet the rule.
Status indicator	Acceptable if the business rule is met. Unacceptable if the business rule is not met.	Unacceptable. Too many orders are not billed.

4.2 Data Profiling

Data Profiling is a form of data analysis used to inspect data and assess quality. Data profiling uses statistical techniques to discover the true structure, content, and quality of a collection of data (Olson, 2003). A profiling engine produces statistics that analysts can use to identify patterns in data content and structure. For example:

- Counts of nulls: Identifies nulls exist and allows for inspection of whether they are allowable or not
- Max/Min value: Identifies outliers, like negatives
- Max/Min length: Identifies outliers or invalids for fields with specific length requirements
- Frequency distribution of values for individual columns: Enables assessment of reasonability (e.g., distribution of country codes for transactions, inspection of frequently or infrequently occurring values, as well as the percentage of the records populated with defaulted values)
- Data type and format: Identifies level of non-conformance to format requirements, as well as identification of unexpected formats (e.g., number of decimals, embedded spaces, sample values)

Profiling also includes cross-column analysis, which can identify overlapping or duplicate columns and expose embedded value dependencies. Inter-table analysis explores overlapping values sets and helps identify foreign key relationships. Most data profiling tools allow for drilling down into the analyzed data for further investigation.

Results from the profiling engine must be assessed by an analyst to determine whether data conforms to rules and other requirements. A good analyst can use profiling results to confirm known relationships and uncover hidden characteristics and patterns within and between data sets, including business rules and validity constraints. Profiling is usually used as part of data discovery for projects (especially data integration projects; see Chapter 8) or to assess the current state of data that is targeted for improvement. Results of data profiling

can be used to identify opportunities to improve the quality of both data and Metadata (Olson, 2003; Maydanchik, 2007).

While profiling is an effective way to understand data, it is just a first step to Data Quality improvement. It enables organizations to identify potential problems. Solving problems requires other forms of analysis, including business process analysis, analysis of data lineage, and deeper data analysis that can help isolate root causes of problems from stakeholders along the data chain. When planning large scale profiling, ensure that time is allocated to share results, prioritize problems, and determine which issues require in-depth analysis.

Create shareable, linkable, and re-usable code modules that execute repeated Data Quality checks and audit processes that developers can get from a library. If the module needs to change, then all the code linked to that module will get updated. Such modules simplify the maintenance process. Well-engineered code blocks can prevent many Data Quality problems. Another benefit is they ensure processes are executed consistently. Where laws or policy mandate reporting of specific quality results, the lineage of results often needs to be described. Quality check modules can provide this. For data that has any questionable quality dimension and that is highly rated, qualify the information in the shared environments with quality notes and confidence ratings.

4.3 Preventive Actions

Create shareable, linkable, and re-usable code modules that execute repeated Data Quality checks and audit processes that developers can get from a library. If the module needs to change, then all the code linked to that module will get updated. Such modules simplify the maintenance process. Well-engineered code blocks can prevent many Data Quality problems. As importantly, they ensure processes are executed consistently. Where laws or policy mandate reporting of specific quality results, the lineage of results often needs to be described. Quality check modules can provide this. For data that has any questionable quality dimension and that is highly rated, qualify the information in the shared environments with quality notes, and confidence ratings.

4.4 Effective Data Quality Metrics

The best way to create high quality data is to prevent poor quality data from entering an organization. Preventive actions stop known errors from occurring. Inspecting data after it is in production will not improve its quality. Approaches include:

- **Establish data entry controls:** Create data entry rules that prevent invalid or inaccurate data from entering a system.
- **Train data producers:** Ensure staff in upstream systems understand the impact of their data on downstream users. Give incentives or base evaluations on data accuracy and completeness, rather than just speed.
- **Define and enforce rules:** Create a ‘data firewall,’ which has a table with all the business Data Quality Rules used to check if the quality of data is good before being used in an application such as a data

warehouse. A data firewall can inspect the level of quality of data processed by an application and if the level of quality is below acceptable levels, analysts can be informed about the problem.

- **Demand high quality data from data suppliers:** Examine an external data provider's processes to check their structures, definitions, data source(s), and data provenance. Doing so enables assessment of how well their data will integrate and helps prevent the use of non-authoritative data or data acquired without permission from the owner.
- **Implement Data Governance and Stewardship:** Ensure roles and responsibilities are defined that describe and enforce rules of engagement, decision rights, and accountabilities for effective management of data and information assets (McGilvray, 2008). Work with data stewards to revise the process of, and mechanisms for, generating, sending, and receiving data.
- **Institute formal change control:** Ensure all changes to stored data are defined and tested before being implemented. Prevent changes directly to data outside of normal processing by establishing gating processes.

4.5 Root Cause Analysis

Data Quality issues should be addressed systemically at their root causes to minimize the costs and risks of corrective actions. 'Solve the problem where it happens' is the best practice in Data Quality Management. A root cause of a problem is a factor that, if eliminated, would remove the problem itself. Root cause analysis is a process of understanding factors that contribute to problems and the ways they contribute. Its purpose is to identify underlying conditions that, if eliminated, would mean problems would disappear. For example, a data process that runs each month requires a file of customer information as input. Measurement of the data shows that in April, July, October and January, the quality of the data goes down. Inspection of the timing of delivery shows that in March, June, September and December, the file is delivered on the 30th of the month, whereas at other times it is delivered on the 25th. Further analysis shows that the team responsible for delivering the file is also responsible for closing quarterly financial processes. These processes take precedence over other work and the files are delivered late during those months, impacting the quality. The root cause of the Data Quality problem turns out to be a process delay caused by a competing priority. It can be addressed by scheduling file delivery and ensuring that resources can deliver within the schedule.

Common techniques for root cause analysis include Pareto analysis (the 80/20 rule), fishbone diagram analysis, track and trace, process analysis, and the Five Whys (Sakichi Toyoda).

4.6 Corrections

Corrections are implemented after a problem has occurred and been detected. Where there have been process, system, or user failures, there is also usually a trail of poor-quality data that needs to be repaired. Corrective actions should include avoiding recurrence of the causes of the quality problems, with process, system changes and user awareness. These fixes should use the activities for upgrading the Data Quality Management in section 2.

There is often a trail of poor-quality data through multiple systems and even into other organizations as modern systems often transmit data in close to real time. Data Cleansing or Scrubbing transforms data to make it conform to data standards and domain rules. Cleansing includes detecting and correcting data errors to bring the quality of data to an acceptable level.

It costs money and introduces risk to continuously remediate data through cleansing. Ideally, the need for data cleansing should decrease over time as root causes of data issues are resolved. In some situations, correcting on an ongoing basis may be necessary, as re-processing the data in a midstream system is cheaper than any other alternative. Perform data cleansing can be performed in three general ways:

- **Fully Automated:** Automated correction techniques include rule-based standardization, normalization and correction. The modified values are obtained or generated and committed without manual intervention. An example is automated address correction which submits delivery addresses to an address standardizer that conforms and corrects delivery addresses using rules, parsing, standardization, and reference tables. Automated correction requires an environment with well-defined standards, commonly accepted rules and known error patterns. The amount of automated correction can be reduced over time if this environment is well-managed and corrected data is shared with upstream systems.
- **Manually-directed:** Use automated tools to remediate and correct data but require manual review before committing the corrections to persistent storage. Apply name and address remediation, identity resolution and pattern-based corrections automatically, and use some scoring mechanism to propose a level of confidence in the correction. Corrections with scores above a particular level of confidence may be committed without review, however, corrections with scores below the level of confidence are presented to the data steward for review and approval. Commit all approved corrections and review those not approved to understand whether to adjust the applied underlying rules. Environments in which sensitive data sets require human oversight (e.g., MDM) are good examples of where manual-directed correction may be suited.
- **Manual:** Sometimes manual correction is the only option in the absence of tools or automation or if it is determined that the change is better handled through human oversight. Manual corrections are best done through an interface with controls and edits, which provide an audit trail for changes. Avoid using the alternative of making corrections and committing the updated records directly in production environments as it is extremely risky.

5. Implementation Guidelines

Improving the quality of data within an organization is not an easy task – even when launched from within a Data Governance Function and with the support of senior management. A classic discussion is whether it is better to implement a Data Quality Function top-down or bottom-up. Typically, a hybrid approach works best – top-down for sponsorship, consistency, and resources, but bottom-up to discover what is actually broken and to achieve incremental successes.

Improving Data Quality requires changes in how people think about and behave toward data. Cultural change is challenging. It requires planning, training, and reinforcement. (See Chapter 17.) While the specifics will differ from organization to organization, most Data Quality Function implementations need to plan for:

- Metrics on the value of high quality data and the cost of poor-quality data: One way to raise organizational awareness of the need for Data Quality Management is through metrics that describe the value of data and the return on investment from improvements. These metrics which differ from Data Quality scores provide the basis for funding improvements and changing the behavior of both staff and management. See Chapter 15.
- Operating model for IT/Business interactions: Business people know what the important data is, and what it means. Data Custodians from IT understand where and how the data is stored and so they are well placed to translate definitions of Data Quality into queries or code that identify specific records that do not comply. See Chapter 12.
- Changes in how projects are executed: Project oversight must ensure project funding includes steps related to Data Quality (e.g., profiling and assessment, definition of quality expectations, data issue remediation, prevention, and correction, building controls and measurements). It is prudent to make sure issues are identified early and to build Data Quality expectations upfront in projects.
- Changes to business processes: Improving Data Quality depends on improving the processes by which data is produced. The Data Quality team needs to be able to assess and recommend changes to non-technical (as well as technical) processes that impact the quality of data.
- Funding for remediation and improvement projects: Some organizations do not plan for remediating data, even when they are aware of Data Quality issues. Data will not fix itself. The costs and benefits of remediation and improvement projects should be documented so that work on improving data can be prioritized.
- Funding for Data Quality Operations: Sustaining Data Quality requires ongoing operations to monitor Data Quality, report on findings and continue to manage issues as they are discovered.

5.1 Readiness Assessment / Risk Assessment

Most organizations that depend on data have a lot of opportunity for improvement. How formal and well-supported a Data Quality function will be depends on how mature the organization is from a data management perspective. See Chapter 15. Organizational readiness to adopt Data Quality practices can be assessed by considering the following characteristics:

- The organization's current understanding of the quality of its data: Before most organizations start their quality improvement journey, they generally have to understand the obstacles and pain points that signify poor quality data. Gaining knowledge of these is important. Through them, poor quality data can be directly associated with negative effects including direct and indirect costs on the organization. An understanding of pain points also helps identify and prioritize improvement projects.

- The actual state of the data: Finding an objective way to describe the condition of data that is causing pain points is the first step to improving the data. Data can be measured and described through profiling and analysis, as well as through quantification of known issues and pain points. If the Data Quality team does not know the actual state of the data, then it will be difficult to prioritize and act on opportunities for improvement.
- Risks associated with data creation, processing, or use: Identifying what can go wrong with data and the potential damage to an organization from poor quality data provides the basis for mitigating risks. If the organization does not recognize these risks, it may be challenging to get support for the Data Quality function.
- Cultural and technical readiness for scalable Data Quality Monitoring: The quality of data can be negatively impacted by business and technical processes. Improving the quality of data depends on cooperation between business and IT teams. If the relationship between business and IT teams is not collaborative, then it will be difficult to make progress.

Findings from a readiness assessment will help determine where to start and how quickly to proceed. Findings can also provide the basis for road-mapping improvements goals. If there is strong support for Data Quality improvement and the organization knows its own data, then it may be possible to launch a full strategic improvement program. If the organization does not know the actual state of its data, then it may be necessary to focus on building that knowledge before developing a full strategy.

5.2 Organization and Cultural Change

The quality of data will not be improved through a collection of tools and concepts. Rather, improvement will come through a mindset that helps employees and stakeholders to act while always thinking of the quality of data and what the business and their customers need. Getting an organization to be conscientious about Data Quality often requires significant cultural change (see Chapter 17). Ultimately, employees need to think and act differently if they are to produce better quality data and manage data in ways that ensure quality. This requires training and reinforcement. Three common themes of how mindsets need to change in Data Quality Management are:

- Lessening the grip of the Status Quo – “the data has always been bad, we have already tried everything to change it.” This attitude destroys the impetus to change and drains team morale.
- Dealing with politics – “they need to do a better job with Data Quality instead of leaving it to me, but I do not have any resources, time, people or assistance to give them.” This attitude comes from siloed natures of organizations and effectively blocks enterprise Data Quality Management.
- Removing hero cultures – “if I did not do this job, no-one else would, I am critical to making this regulatory report happen”. This attitude ensures opaqueness in Data Quality measurement, avoiding objective measurement and transparency (key Data Quality principle).

The first step is promoting awareness about the role and importance of data to the organization. All employees must act responsibly and raise Data Quality issues, ask for good quality data as consumers, and provide quality

information to others. Every person who touches the data can impact the quality of that data. Data Quality is not just the responsibility of a Data Quality team or IT group.

Just as the employees need to understand the cost to acquire a new customer or retain an existing customer, they also need to know the organizational costs of poor-quality data, as well as the conditions that cause data to be of poor quality. For example, if customer data is incomplete, a customer may receive the wrong product, creating direct and indirect costs to an organization. Not only will the customer return the product, but he or she may call and complain, using call center time, with the potential for reputational damage to the organization. If customer data is incomplete because the organization has not established clear requirements, then everyone who uses this data has a stake in clarifying requirements and following standards.

6. Data Quality and the other Knowledge Areas

All data management knowledge areas contribute to the quality of data. High quality data that supports the organization should be the goal of all data management disciplines. Cross-functional commitment and coordination produces high quality data. Uninformed decisions or actions by anyone who interacts with data can result in poor quality data. Organizations and teams should be aware of this and should plan for high quality data by executing processes and projects in ways that account for risk related to unexpected or unacceptable conditions in the data. This section identifies the key relationships that Data Quality has with other DMBOK Knowledge Areas.

6.1 Data Quality and Data Modeling and Design

Data Quality is enabled by Data Modelling and Design during the design phase of projects by allowing for data elements to:

- Hold Data Quality values. For example, after a customer address is validated, then the address may be marked as passing validation. This is an additional data element that needs to be accounted for in the design
- Associate business rules. For example, if a customer address must always be completed to a certain level, then it is more efficient to store the rule in the data model against the element than in every business process that uses and updates the address.

Tools used to model data and create ETL processes have a direct impact on the quality of data. If they are used without knowledge of the data, they can have detrimental effects. The Data Quality team should work with development teams to ensure that Data Quality risks are addressed and that the organization takes full advantage of the ways in which effective modeling and data processing can enable higher quality data. (See Chapters 5, 8, and 11.)

6.2 Data Quality and Metadata Management

Metadata defines what the data represents. Having a robust process by which data is defined supports the ability of an organization to formalize and document the standards and requirements by which the quality of data can be measured. Data Quality is about meeting expectations. Metadata is a primary means of clarifying expectations.

The most important way that Data Quality Management and Metadata Management are related is that Metadata is created during the Data Quality process. This Metadata should remain associated with the data. For example, if a customer address is marked as valid, complete and reasonable by a process, then this needs to be maintained as Metadata with the address. It assists the Data Quality principles of being standards driven with objective measurement and transparency.

Metadata holds to express the “fitness” of the data. This enables consumers to assess the quality of data based on how well it meets the “purpose” of particular data consumer contexts. Conversely, if there is no Data Quality Metadata associated with data then users will not trust the data resulting in significant effort spent rechecking and validation. Data Quality teams should work closely with teams that manage Metadata to ensure that Data Quality requirements, rules, measurement results and documentation of issues are made available to data consumers. A Metadata repository can house results of Data Quality measurements so that these are shared across the organization and the Data Quality team can work toward consensus about priorities and drivers for improvement. (See Chapter 12.)

6.3 Data Quality and Master and Reference Data Management

Reference and Master Data Management controls the copying of shared data across the organization. Shared data is used in many events to control the flow of business processes, enable decision making and validate data. Therefore, most (if not all) master and reference data is critical data within the Data Quality regime.

A key source of poor Data Quality is Weak Master Data Management: Immature Master Data Management can lead to choosing unreliable sources for data which can cause Data Quality issues that are very difficult to find until the assumption that the data source is accurate is disproved. Reference and master data is a key source of data for Data Quality rule enforcement, particularly the Data Quality dimensions of validity, consistency and integrity. For example, a text value must be in a defined domain of values. These values will be managed Reference Data. For example, in order to have integrity, an order must be on an existing, approved customer. The customer record is master data that must be managed to be ready and available to Data Quality validation processing.

6.4 Data Quality and Data Integration and Interoperability

Data Quality is related to data integration and interoperability as the movement of data is a common source of errors within data processes. During data movement it is important that critical Metadata is maintained with the data (and integrations do not just remove the Metadata). Data Quality Metadata provides trust and confidence in the data with assurance that it is fit for defined purposes.

The other key value of data integrations is that is a key place to test the Data Quality. By monitoring Data Quality between processing stages (i.e., at key integration points), data that is corrupted can quickly be identified and isolated, avoiding propagation of data errors through full data value chains.

6.5 Data Quality and Data Governance

A Data Quality Function is more effective when part of a Data Governance Function. Often Data Quality issues are the reason for establishing enterprise-wide Data Governance (see Chapter 3). A Governance Organization can accelerate the work of a Data Quality Function by:

- Setting priorities and providing direction
- Providing mechanisms to facilitate staff involvement and knowledge sharing
- Identifying and coordinating access to those who should be involved in Data Quality decisions and activities
- Ensuring measurements of Data Quality are acted upon
- Sharing Data Quality inspection results, identify opportunities for improvement, and build consensus
- Resolving variations and conflicts

Incorporating Data Quality efforts into the overall governance effort enables the Data Quality Function team to work with a range of stakeholders and enables:

- Business process engineering and training staff who can help teams implement process improvements
- Business and operational data stewards, and data owners who can identify critical data, define standards and quality expectations, and prioritize remediation of data issues.

7. Works Cited / Recommended

- Batini, Carlo, and Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Springer, 2006. Print.
- Brackett, Michael H. *Data Resource Quality: Turning Bad Habits into Good Practices*. Addison-Wesley, 2000. Print.
- Deming, W. Edwards. *Out of the Crisis*. The MIT Press, 2000. Print.
- English, Larry. *Improving Data Warehouse and Business Information Quality: Methods For Reducing Costs And Increasing Profits*. John Wiley and Sons, 1999. Print.
- English, Larry. *Information Quality Applied: Best Practices for Improving Business Information, Processes, and Systems*. Wiley Publishing, 2009. Print.
- Evans, Nina and Price, James. "Barriers to the Effective Deployment of Information Assets: An Executive Management Perspective." *Interdisciplinary Journal of Information, Knowledge, and Management* Volume 7, 2012. Accessed from <http://bit.ly/2sVwvG4>.
- Fisher, Craig, Eitel Lauría, Shobha Chengalur-Smith and Richard Wang. *Introduction to Information Quality*. M.I.T. Information Quality Program Publications, 2006. Print. *Advances in Information Quality Book Ser.*
- Gottesdiener, Ellen. *Requirements by Collaboration: Workshops for Defining Needs*. Addison-Wesley Professional, 2002. Print.
- Hass, Kathleen B. and Rosemary Hossenlopp. *Unearthing Business Requirements: Elicitation Tools and Techniques*. Management Concepts, Inc, 2007. Print. *Business Analysis Essential Library*.
- Huang, Kuan-Tsae, Yang W. Lee and Richard Y. Wang. *Quality Information and Knowledge*. Prentice Hall, 1999. Print.
- Jugulum, Rajesh. *Competing with High Quality Data*. Wiley, 2014. Print.
- Lee, Yang W., Leo L. Pipino, James D. Funk and Richard Y. Wang. *Journey to Data Quality*. The MIT Press, 2006. Print.

- Loshin, David. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann, 2001. Print.
- Loshin, David. *Master Data Management*. Morgan Kaufmann, 2009. Print.
- Maydanchik, Arkady. *Data Quality Assessment*. Technics Publications, LLC, 2007 Print.
- McCallum, Ethan. *Bad Data Handbook: Cleaning Up the Data So You Can Get Back to Work*. 1st Edition. O'Reilly, 2012.
- McGilvray, Danette. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann, 2008. Print.
- Myers, Dan. "The Value of Using the Dimensions of Data Quality", *Information Management*, August 2013. <http://bit.ly/2tsMYiA>.
- Olson, Jack E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003. Print.
- Redman, Thomas. *Data Quality: The Field Guide*. Digital Press, 2001. Print.
- Robertson, Suzanne and James Robertson. *Mastering the Requirements Process: Getting Requirements Right*. 3rd ed. Addison-Wesley Professional, 2012. Print.
- Sebastian-Coleman, Laura. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Morgan Kaufmann, 2013. Print. The Morgan Kaufmann Series on Business Intelligence.
- Tavares, Rossano. *Qualidade de Dados em Gerenciamento de Clientes (CRM) e Tecnologia da Informação [Data Quality in Management of Customers and Information Technology]*. São Paulo: Catálise. 2006. Print.
- Witt, Graham. *Writing Effective Business Rules: A Practical Method*. Morgan Kaufmann, 2012. Print.

8. Appendix

8.1 Data Quality ISO Standard

ISO 8000, the international standard for Data Quality, is being developed to enable the exchange of complex data in an application-neutral form. In the introduction to the standard, ISO asserts: "The ability to create, collect, store, maintain, transfer, process and present data to support business processes in a timely and cost-effective manner requires both an understanding of the characteristics of the data that determine its quality, and an ability to measure, manage and report on Data Quality."

ISO 8000 defines dimensions that can be tested by any organization in the data supply chain to objectively determine conformance of the data to ISO 8000.⁷⁶

The first published part of ISO 8000 (part 110, published in 2008) focused on the syntax, semantic encoding, and conformance to the data specification of Master Data. Other parts projected for the standard include part 100 - Introduction, part 120 - Provenance, part 130 -Accuracy, and part 140 - Completeness.⁷⁷

ISO 8000 defines quality data as "portable data that meets stated requirements."⁷⁸ The Data Quality standard is related to the ISO's overall work on data portability and preservation. Data is considered 'portable' if it can be separated from a software application. Data that can only be used or read using a specific licensed software

⁷⁶ <http://bit.ly/2ttdiZJ>.

⁷⁷ <http://bit.ly/2sANGdi>.

⁷⁸ <http://bit.ly/2rV1oWC>.

application is subject to the terms of the software license. An organization may not be able to use data it created unless that data can be detached from the software that was used to create it.

To meet stated requirements requires that these requirements be defined in a clear, unambiguous manner. ISO 8000 is supported through ISO 22745, a standard for defining and exchanging Master Data. ISO 22745 defines how data requirement statements should be constructed, provides examples in XML, and defines a format for the exchange of encoded data.⁸¹ ISO 22745 creates portable data by labeling the data using an ISO 22745 compliant Open Technical Dictionary such as the ECCMA Open Technical Dictionary (eOTD).

The intention of ISO 8000 is to help organizations define what is and is not quality data, enable them to ask for quality data using standard conventions, and verify that they have received quality data using those same standards. When standards are followed, requirements can be confirmed through a computer program.

ISO 8000 - Part 61 Information and Data Quality Management process reference model is under development. This standard will describe the structure and organization of Data Quality Management, including:

- Data Quality Planning
- Data Quality Control
- Data Quality Assurance
- Data Quality Improvement

8.2 Further Reading on Data Quality Dimensions

Some dimensions can be measured objectively (completeness, validity, format conformity) and others that depend on heavily context or on subjective interpretation (usability, reliability, reputation). Whatever names are used, dimensions focus on whether there is enough data (completeness), whether it is right (accuracy, validity), how well it fits together (consistency, integrity, uniqueness), whether it is up-to-date (timeliness), accessible, usable, and secure.

Many leading thinkers in Data Quality have published sets of dimensions.⁷⁹ The three most influential are described here because they provide insight into how to think about what it means to have high quality data, as well as into how Data Quality can be measured.

The Strong-Wang framework (1996) focuses on data consumers' perceptions of data. It describes 15 dimensions across four general categories of Data Quality:

- Intrinsic Data Quality
 - Accuracy
 - Objectivity
 - Believability
 - Reputation
 - Contextual Data Quality
-

⁷⁹ In addition to the examples detailed here and numerous academic papers on this topic, see Loshin (2001), Olson (2003), McGilvray (2008), and Sebastian-Coleman (2013) for detailed discussions on Data Quality dimensions. See Myers (2013) for a comparison of dimensions.

- Value-added
- Relevancy
- Timeliness
- Completeness
- Appropriate amount of data
- Representational Data Quality
 - Interpretability
 - Ease of understanding
 - Representational consistency
 - Concise representation
- Accessibility Data Quality
 - Accessibility
 - Access security

In *Data Quality for the Information Age* (1996), Thomas Redman formulated a set of Data Quality dimensions rooted in data structure.⁸⁰ Redman defines a data item as a “representable triple”: a value from the domain of a data element within an entity. Dimensions can be associated with any of the component pieces of data – the model (entities and data elements) as well as the values. Redman includes the dimension of representation, which he defines as a set of rules for recording data items. Within these three general categories (data model, data values, representation), he describes more than two dozen dimensions. They include the following:

Data Model:

- Content:
 - Relevance of data
 - The ability to obtain the values
 - Clarity of definitions
- Level of detail:
 - Data element granularity
 - Precision of data element domains
- Composition:
 - Naturalness: The idea that each data element should have a simple counterpart in the real world and that each data element should bear on a single fact about the entity
 - Identify-ability: Each entity should be distinguishable from every other entity
 - Homogeneity
 - Minimum necessary redundancy
- Consistency:
 - Semantic consistency of the components of the model
 - Structure consistency of data elements across entity types
- Reaction to change:
 - Robustness
 - Flexibility

⁸⁰ Redman expanded and revised his set of dimensions in *Data Quality: The Field Guide* (2001).

Data Values:

- Accuracy
- Completeness
- Currency
- Consistency

Representation:

- Appropriateness
- Interpretability
- Portability
- Format precision
- Format flexibility
- Ability to represent null values
- Efficient use of storage
- Physical instances of data being in accord with their formats

Redman recognizes that consistency of entities, values, and representation can be understood in terms of constraints. Different types of consistency are subject to different kinds of constraints.

In *Improving Data Warehouse and Business Information Quality* (1999), Larry English presents a comprehensive set of dimensions divided into two broad categories: inherent and pragmatic.⁸¹ Inherent characteristics are independent of data use. Pragmatic characteristics are associated with data presentation and are dynamic; their value (quality) can change depending on the uses of data.

- **Inherent** quality characteristics
 - Definitional conformance
 - Completeness of values
 - Validity or business rule conformance
 - Accuracy to a surrogate source
 - Accuracy to reality
 - Precision
 - Non-duplication
 - Equivalence of redundant or distributed data
 - Concurrency of redundant or distributed data
- **Pragmatic** quality characteristics
 - Accessibility
 - Timeliness
 - Contextual clarity
 - Usability
 - Derivation integrity
 - Rightness or fact completeness

⁸¹ English expanded and revised his dimensions in *Information Quality Applied* (2009).

In 2013, DAMA UK produced a white paper describing six core dimensions of Data Quality:

- **Completeness:** The proportion of data stored against the potential for 100%.
- **Uniqueness:** No entity instance (thing) will be recorded more than once based upon how that thing is identified.
- **Timeliness:** The degree to which data represent reality from the required point in time.
- **Validity:** Data is valid if it conforms to the syntax (format, type, range) of its definition.
- **Accuracy:** The degree to which data correctly describes the ‘real world’ object or event being described.
- **Consistency:** The absence of difference, when comparing two or more representations of a thing against a definition.

The DAMA UK white paper also describes other characteristics that have an impact on quality. While the white paper does not call these dimensions, they work in a manner similar to Strong and Wang’s contextual and representational Data Quality and English’s pragmatic characteristics.

- **Usability:** Is the data understandable, simple, relevant, accessible, maintainable and at the right level of precision?
- **Timing issues** (beyond timeliness itself): Is it stable yet responsive to legitimate change requests?
- **Flexibility:** Is the data comparable and compatible with other data? Does it have useful groupings and classifications? Can it be repurposed? Is it easy to manipulate?
- **Confidence:** Are Data Governance, Data Protection, and Data Security processes in place? What is the reputation of the data, and is it verified or verifiable?
- **Value:** Is there a good cost / benefit case for the data? Is it being optimally used? Does it endanger people’s safety or privacy, or the legal responsibilities of the enterprise? Does it support or contradict the corporate image or the corporate message?

8.3 Statistical Process Control

Statistical Process Control (SPC) is a method to manage processes by analyzing measurements of variation in process inputs, outputs, or steps. The technique was developed in the manufacturing sector in the 1920s and has been applied in other industries, in improvement methodologies such as Six Sigma, and in Data Quality Management. Simply defined, a process is a series of steps executed to turn inputs into outputs. SPC is based on the assumption that when a process with consistent inputs is executed consistently, it will produce consistent outputs. It uses measures of central tendency (how values cluster around a central value, such as a mean, median, or mode) and of variability around a central value (e.g., range, variance, standard deviation), to establish tolerances for variation within a process. The primary tool used for SPC is the control chart (Figure 95), which is a time series graph that includes a central line for the average (the measure of central tendency) and depicts calculated upper and lower control limits (variability around a central value). In a stable process, measurement results outside the control limits indicate a special cause.

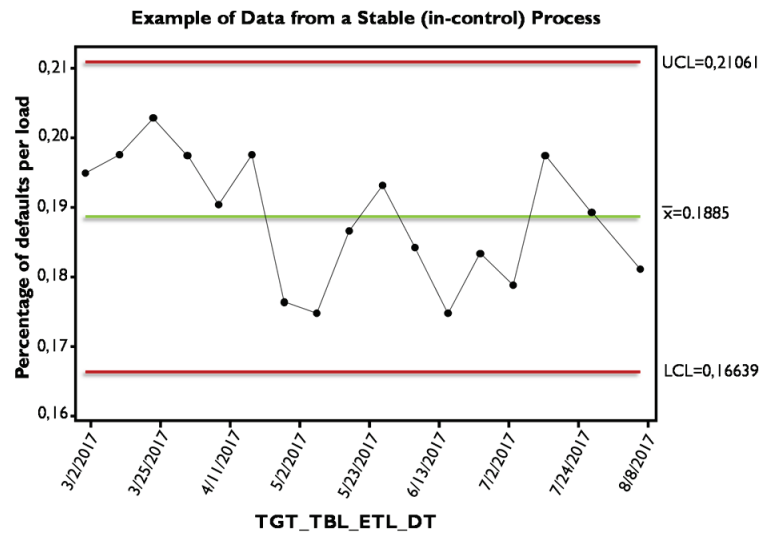


Figure 95 Control Chart of a Process in Statistical Control

SPC measures the predictability of process outcomes by identifying variation within a process. Processes have variation of two types: Common Causes that are inherent in the process and Special Causes that are unpredictable or intermittent. When the only sources of variation are common causes, a system is said to be in (statistical) control and a range of normal variation can be established. This is the baseline against which change can be detected.

Applying SPC to Data Quality measurement is based on the working assumption that, like a manufactured product, data is the product of a process. Sometimes the process that creates data is very simple (e.g., a person fills out a form). Other times, processes are quite complex: a set of algorithms aggregates medical claim data in order to follow trends related to the effectiveness of particular clinical protocols. If such a process has consistent inputs and is executed consistently, it will produce consistent results each time it is run. However, if the inputs or execution change, then so will the outputs. Each of these components can be measured. The measurements can be used to detect special causes. Knowledge of the special causes can be used to mitigate risks associated with data collection or processing.

SPC is used for control, detection, and improvement. The first step is to measure the process in order to identify and eliminate special causes. This activity establishes the control state of the process. Next is to put in place measurements to detect unexpected variation as soon as it is detectable. Early detection of problems simplifies investigation of their root causes. Measurements of the process can also be used to reduce the unwanted effects of common causes of variation, allowing for increased efficiency.