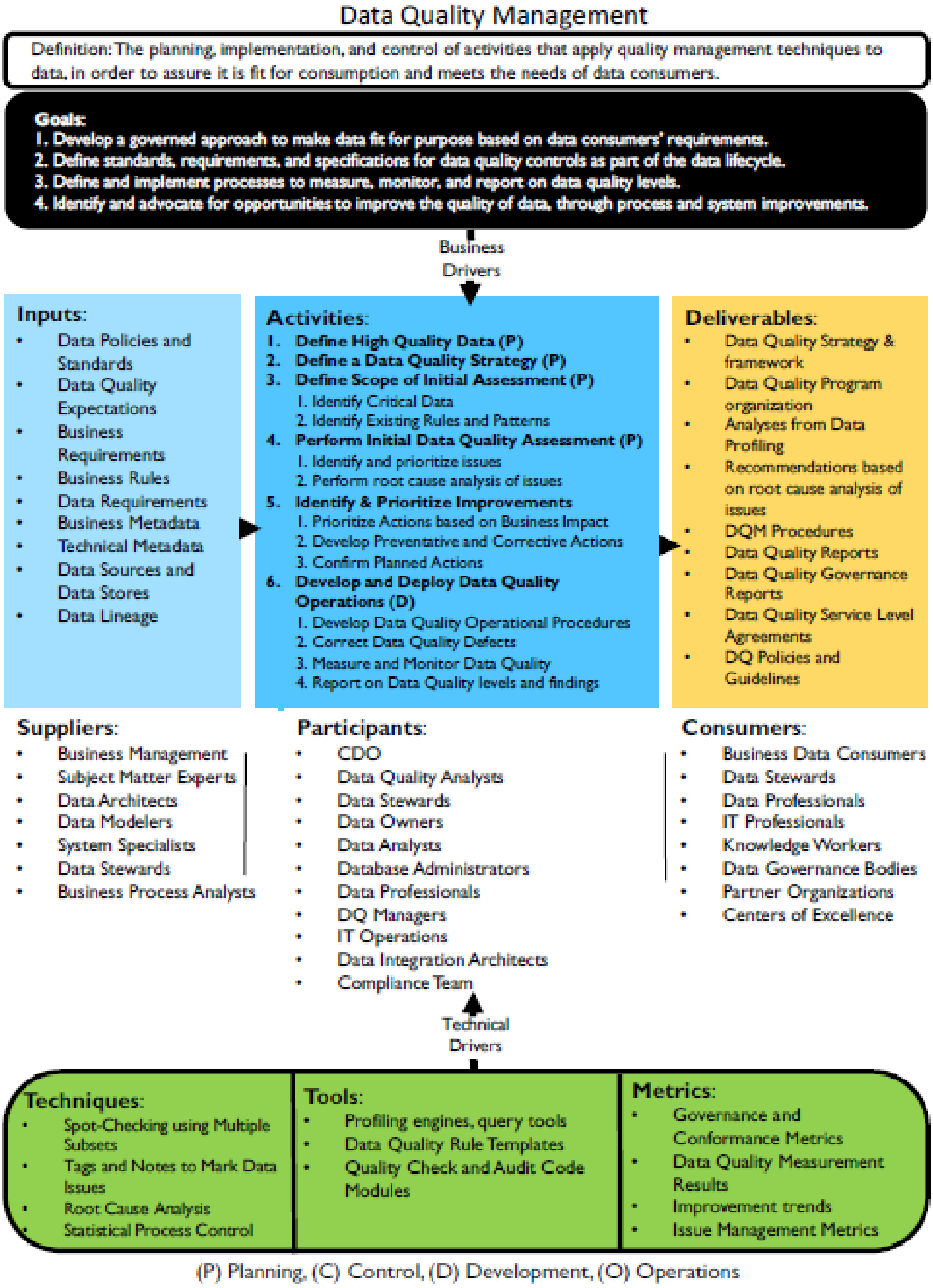
Data Quality

1 Introduction



Data Quality Management

Definition: 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.

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
- Data Sources and Data Stores
- Data Lineage

Activities:
1. Define High Quality Data (P)
2. Define a Data Quality Strategy (P)
3. Define Scope of Initial Assessment (P)
   1. Identify Critical Data
   2. Identify Existing Rules and Patterns
4. Perform Initial Data Quality Assessment (P)
   1. Identify and prioritize issues
   2. Perform root cause analysis of issues
5. Identify & Prioritize Improvements
   1. Prioritize Actions based on Business Impact
   2. Develop Preventative and Corrective Actions
   3. Confirm Planned Actions
6. Develop and Deploy Data Quality Operations (D)
   1. Develop Data Quality Operational Procedures
   2. Correct Data Quality Defects
   3. Measure and Monitor Data Quality
   4. Report on Data Quality levels and findings

Deliverables:
- Data Quality Strategy & framework
- Data Quality Program organization
- Analyses from Data Profiling
- Recommendations based on root cause analysis of issues
- DQM Procedures
- Data Quality Reports
- Data Quality Governance Reports
- Data Quality Service Level Agreements
- DQ Policies and Guidelines

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

Participants:
- CDO
- Data Quality Analysts
- Data Stewards
- Data Owners
- Data Analysts
- Database Administrators
- Data Professionals
- DQ Managers
- IT Operations
- Data Integration Architects
- Compliance Team

Consumers:
- Business Data Consumers
- Data Stewards
- Data Professionals
- IT Professionals
- Knowledge Workers
- Data Governance Bodies
- Partner Organizations
- Centers of Excellence

Technical Drivers

Techniques:
- Spot-Checking using Multiple Subsets
- Tags and Notes to Mark Data Issues
- Root Cause Analysis
- Statistical Process Control

Tools:
- Profiling engines, query tools
- Data Quality Rule Templates
- Quality Check and Audit Code Modules

Metrics:
- Governance and Conformance Metrics
- Data Quality Measurement Results
- Improvement trends
- Issue Management Metrics

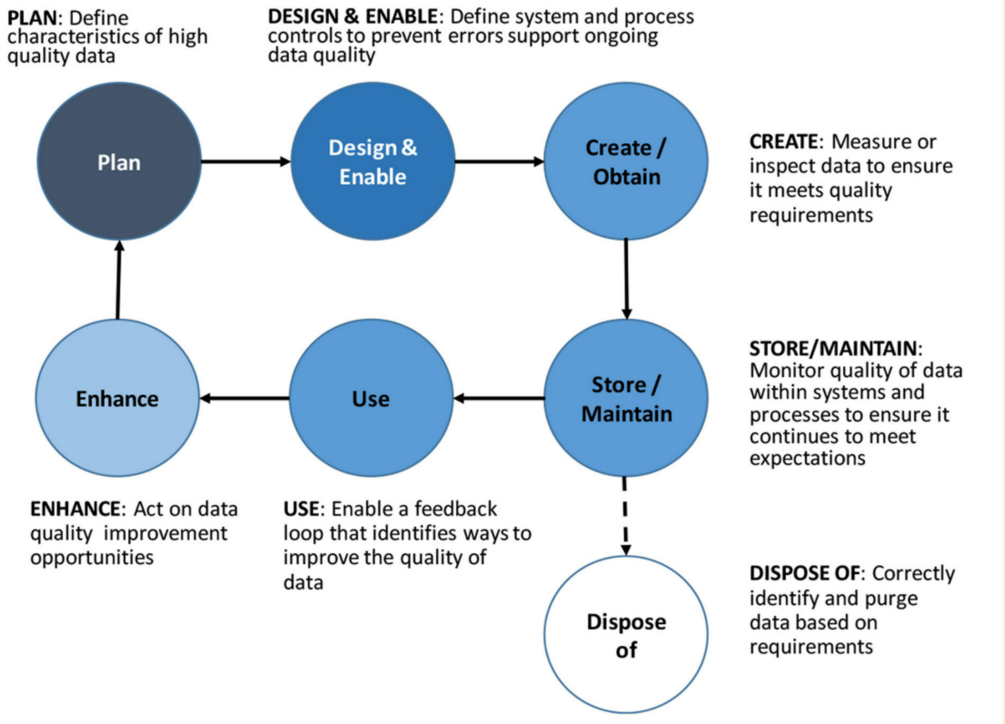
(P) Planning, (C) Control, (D) Development, (O) Operations

The value of data is that data is reliable and trustworthy i.e. of high quality.

Factors that contribute to poor quality data:

- Lack of understanding of the effects of poor quality data on the decision-making process
- Bad Planning
- Siloed system design
- Inconsistent development processes
- Incomplete documentation
- A lack of standards
- A lack of governance
- Failure to define what makes data fit for purpose

High quality data should be the goal of all data management disciplines and they all contribute to the quality of data. Data Quality should be managed by a Data Quality Program team as Data Quality is an enterprise program like Data Governance.



PLAN: Define characteristics of high quality data

DESIGN & ENABLE: Define system and process controls to prevent errors support ongoing data quality

CREATE: Measure or inspect data to ensure it meets quality requirements

STORE/MAINTAIN: Monitor quality of data within systems and processes to ensure it continues to meet expectations

USE: Enable a feedback loop that identifies ways to improve the quality of data

ENHANCE: Act on data quality improvement opportunities

DISPOSE OF: Correctly identify and purge data based on requirements

Figure 27: Data Quality Management and the Data Lifecycle (Adapted from DMBOK2, p. 29)

From Navigating the Labrynth

1.1 Business Drivers

Business drivers for establishing a Data Quality Program:

- Increase value of business data and opportunities to use it
- Reduce risks associated with poor quality data
- Improve organisational efficiency and productivity
- Protect and enhance the organisation's reputation

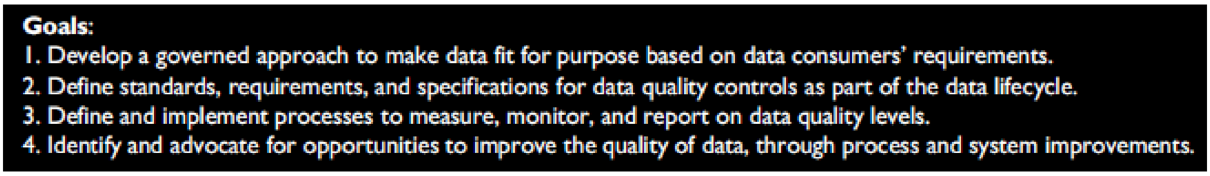
Direct costs associated with poor quality data include:

- Inability to invoice correctly

- Increased customer service calls and decreased ability to resolve them
- Revenue lost due to missed opportunities
- Delay if integration during mergers or acquisitions
- increased exposure to fraud
- loss due to bad business decisions driven by bad data
- Loss of business due to poor credit standing

1.2 Goals and Principles

General goals:



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.

A Data Quality Program should be guided by the following principles:

- **Criticality:** Focus on data most critical to the enterprise and its customers
- **Lifecycle management:** Manage data across data lifecycle from creation through disposal.
- **Prevention:** Focus should be on prevention of data errors
- **Root cause remediation:** Don't just correct errors, address the root cause
- **Governance:** Support from Data Governance activities
- **Standards-driven**: Define requirements as measurable standards
- **Objective measurement and transparency:** Measurement and methodology must be communicated to stakeholders
- **Embedded in business processes:** Business process owners must enforce data quality standards
- **Systematically enforced:** System owners must enforce data quality requirements
- **Connected to service levels:** Data quality reporting and issues management must be incorporated into SLAs.

1.3 Essential Concepts

1.3.1 Data Quality

Data is of high quality when it meets the expectations and needs of data consumers, i.e. it is fit for the purpose to which they want to apply it.

1.3.2 Critical Data

How to identify critical data. It is usually required by:

- Regulatory reporting
- Financial reporting
- Business policy
- Ongoing operations
- Business strategy, especially efforts at competitive differentiation
- MASTERDATA is always critical

1.3.3 Data Quality Dimensions

A measurable characteristic of data which form the basis for measurable rules.

DAMA UK white paper (2013) – six core dimensions:

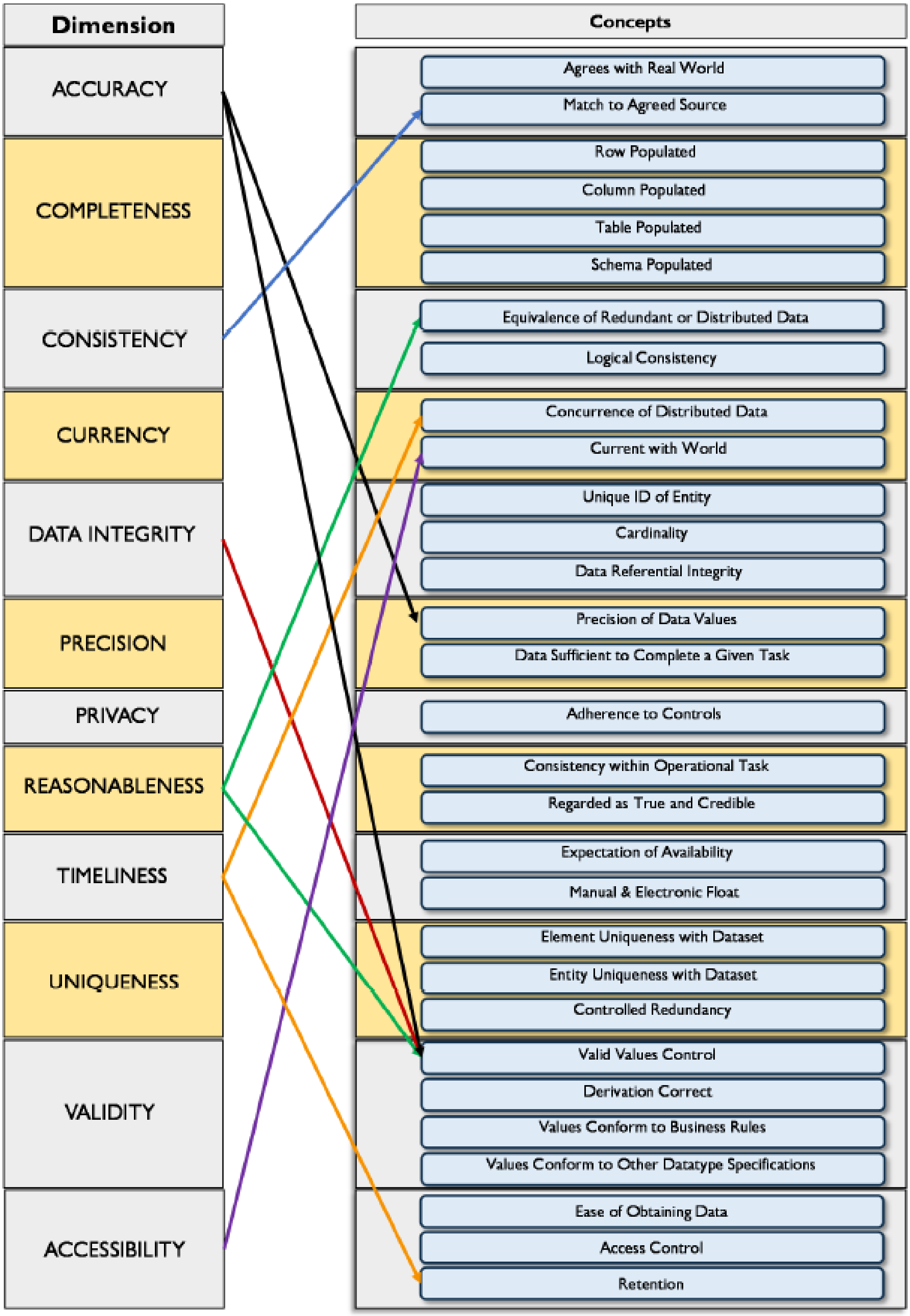
- **Completeness:** Proportion of data stored against 100%
- **Uniqueness:** no entity instance recorded more than once
- **Timeliness:** Degree to which data represents reality at any point in time
- **Validity:** Data conforms to the syntax of its definition
- **Accuracy:** degree to which data describes the real world
- **Consistency:** no difference found when comparing two or more representations of a thing to definitions

Other useful measurable characteristic (not listed as dimensions):

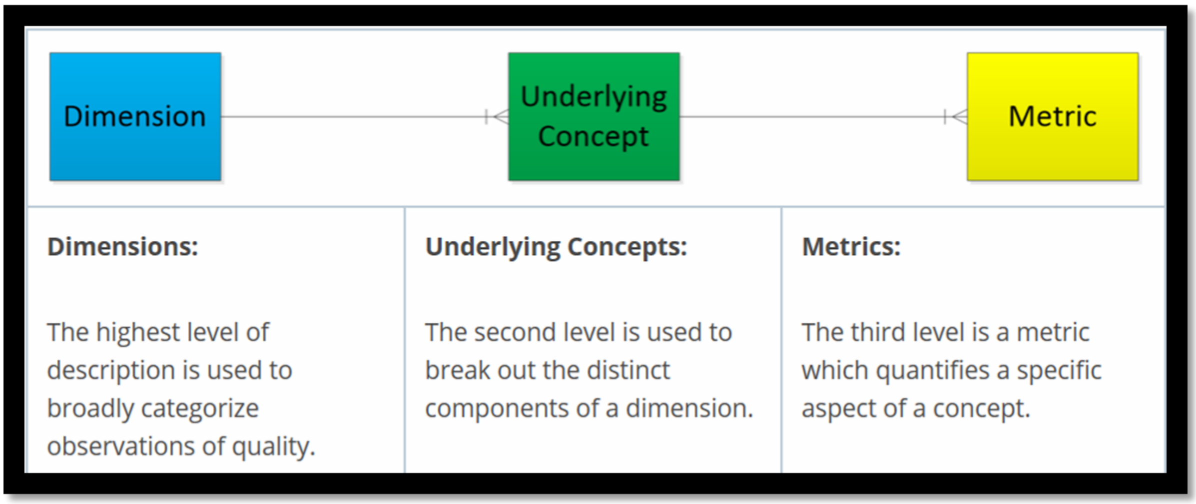
- **Usability:** Is the data understandable, simple, relevant, accessible, maintainable, and at the right level of precision?
- **Timing issues:** Is it stable yet responsive to legitimate change requests?
- **Flexibility:** Can it be repurposed? Is it easy to manipulate?
- **Confidence:** Are Data Governance processes in place? Is the data verified?
- **Value:** Good benefit case for the data? Is it being optimally used?

Dimension of Quality	Description
Accuracy	Accuracy refers to the degree that data correctly represents 'real-life' entities. Accuracy is difficult to measure, unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source (e.g., Dun and Bradstreet Reference Data).
Completeness	Completeness refers to whether all required data is present. Completeness can be measured at the data set, record, or column level. Does the data set contain all the records expected? Are records populated correctly? (Records with different statuses may have different expectations for completeness.) Are columns/attributes populated to the level expected? (Some columns are mandatory. Optional columns are populated only under specific conditions.) Assign completeness rules to a data set with varying levels of constraint: Mandatory attributes that require a value, data elements with conditional and optional values, and inapplicable attribute values. Data set level measurements may require comparison to a source of record or may be based on historical levels of population.
Consistency	Consistency can refer to ensuring that data values are consistently represented within a data set and between data sets, and consistently associated across data sets. It can also refer to the size and composition of data sets between systems or across time. Consistency may be defined between one set of attribute values and another attribute set within the same record (record-level consistency), between one set of attribute values and another attribute set in different records (cross-record consistency), or between one set of attribute values and the same attribute 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. Characteristics that are expected to be consistent within and across data sets can be used as the basis for standardizing data. Data Standardization refers to the conditioning of input data to ensure that data meets rules for content and format. Standardizing data enables more effective matching and facilitates consistent output. Encapsulate consistency constraints as a set of rules that specify consistent relationships between values of attributes, either across a record or message, or along all values of a single attribute (such as a range or list of valid values). For example, one might expect that the number of transactions each day does not exceed 105% of the running average number of transactions for the previous 30 days.
Integrity	Data Integrity (or Coherence) includes ideas associated with completeness, accuracy, and consistency. In data, integrity usually refers to either referential integrity (consistency between data objects via a reference key contained in both objects) or internal consistency within a data set such that there are no holes or missing parts. 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. The level of orphan records can be measured as a raw count or as a percentage of the data set.
Reasonability	Reasonability asks whether a data pattern meets expectations. For example, whether a distribution of sales across a geographic area makes sense based on what is known about the customers in that area. Measurement of reasonability can take different forms. For example, reasonability may be based on comparison to benchmark data, or past instances of a similar data set (e.g., sales from the previous quarter). Some ideas about reasonability may be perceived as subjective. If this is the case, work with data consumers to articulate the basis of their expectations of data to formulate objective comparisons. Once benchmark measurements of reasonability are established, these can be used to objectively compare new instances of the same data set in order to detect change. (See Section 4.5.)
Timeliness	The concept of data Timeliness refers to several characteristics of data. Measures of timeliness need to be understood in terms of expected volatility – how frequently data is likely to change and for what reasons. Data currency is the measure of whether data values are the most up-to-date version of the information. Relatively static data, for example some Reference Data values like country codes, may remain current for a long period. Volatile data remains current for a short period. Some data, for example, 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. Once markets close, the data will remain unchanged, but will still be current, since the market itself is inactive. Latency measures the time between when the data was created and when it was made available for use. For example, overnight batch processing can give a latency of 1 day at 8am for data entered into the system during the prior day, but only one hour for data generated during the batch processing. (See Chapter 8.)
Uniqueness / Deduplication	Uniqueness states that no entity exists more than once within the data set. Asserting uniqueness of the entities within a data set implies that a key value relates to each unique entity, and only that specific entity, within the data set. Measure uniqueness by testing against key structure. (See Chapter 5.)
Validity	Validity refers to whether data values are consistent with a defined domain of values. A domain of values may be a defined set of valid values (such as in a reference table), a range of values, or value that can be determined via rules. The data type, format, and precision of expected values must be accounted for in defining the domain. Data 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. Validate data by comparing it to domain constraints. Keep in mind that data may be valid (i.e., it may meet domain requirements) and still not be accurate or correctly associated with particular records.

Relationship between Data Quality Dimensions and Data Quality Concepts:



Dimension	Concepts
ACCURACY	Agrees with Real World Match to Agreed Source
COMPLETENESS	Row Populated Column Populated Table Populated Schema Populated
CONSISTENCY	Equivalence of Redundant or Distributed Data Logical Consistency
CURRENCY	Concurrence of Distributed Data Current with World
DATA INTEGRITY	Unique ID of Entity Cardinality Data Referential Integrity
PRECISION	Precision of Data Values Data Sufficient to Complete a Given Task
PRIVACY	Adherence to Controls
REASONABLENESS	Consistency within Operational Task Regarded as True and Credible
TIMELINESS	Expectation of Availability Manual & Electronic Float
UNIQUENESS	Element Uniqueness with Dataset Entity Uniqueness with Dataset Controlled Redundancy
VALIDITY	Valid Values Control Derivation Correct Values Conform to Business Rules Values Conform to Other Datatype Specifications
ACCESSIBILITY	Ease of Obtaining Data Access Control Retention



Dimensions:	Underlying Concepts:	Metrics:
The highest level of description is used to broadly categorize observations of quality.	The second level is used to break out the distinct components of a dimension.	The third level is a metric which quantifies a specific aspect of a concept.

1.4 Data Quality and Metadata

Metadata defines what data represents. A Metadata repository can house the results of data quality measurements so that they can be shared, and expectations clarified.

1.5 Data Quality ISO Standard

ISO 8000 is the international standard for data quality. ISO 8000 defines the characteristics that can be tested by any organisation in the data supply chain to objectively determine conformance.

ISO 8000 defines quality data as "portable data that meets stated requirements". Portable means that the data can be separated from a software application. Stated requirements must be clearly defined.

1.6 Data Quality Improvement Lifecycle

Approach data quality improvement based on the technique of quality improvement in physical products. Outputs from one process become inputs to other processes and can impact data quality. Use "plan-do-check-act" from problem solving model, the Shewhart/Deming Cycle.

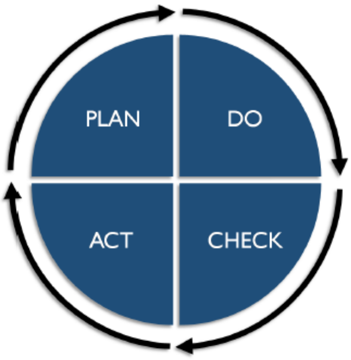


Figure 93 The Shewhart Chart

- **Plan Stage**: DQ team assesses scope, impact and priority of known issues, and evaluates alternatives to address them at the root cause.
- **Do stage**: DQ team leads efforts to address issues at the root cause and plans for ongoing monitoring of data
- **Check stage**: Actively monitor data against standards. If data falls below, act to bring it to acceptable levels
- **Act stage**: Activities to address emerging data issues. The cycle restarts.

1.7 Data Quality Business Rule Types

Data Quality Business Rules describe how data should exist in order to be useful within the organisation. Implemented in software or data entry templates.

Common business rule types:

- **Definitional conformance:** Data definitions used properly across organisation
- **Value presence and record completeness:** Rules for acceptability of missing values
- **Format compliance:** Values have a pattern e.g. phone numbers
- **Value domain membership:** exists in a defined data domain. (Master and reference data)
- **Range conformance:** within a defined range of values
- **Mapping conformance:** Maps to a domain (Reference data)
- **Consistency rules:** Maintain a relationship between two attributes based on the values
- **Accuracy verification:** Compare value to trusted source
- **Uniqueness verification:** Specify which entities must have unique representation. Primary key.
- **Timeliness validation:** Characteristics associated with expectations for accessibility and availability of data, and also when it was last updated.

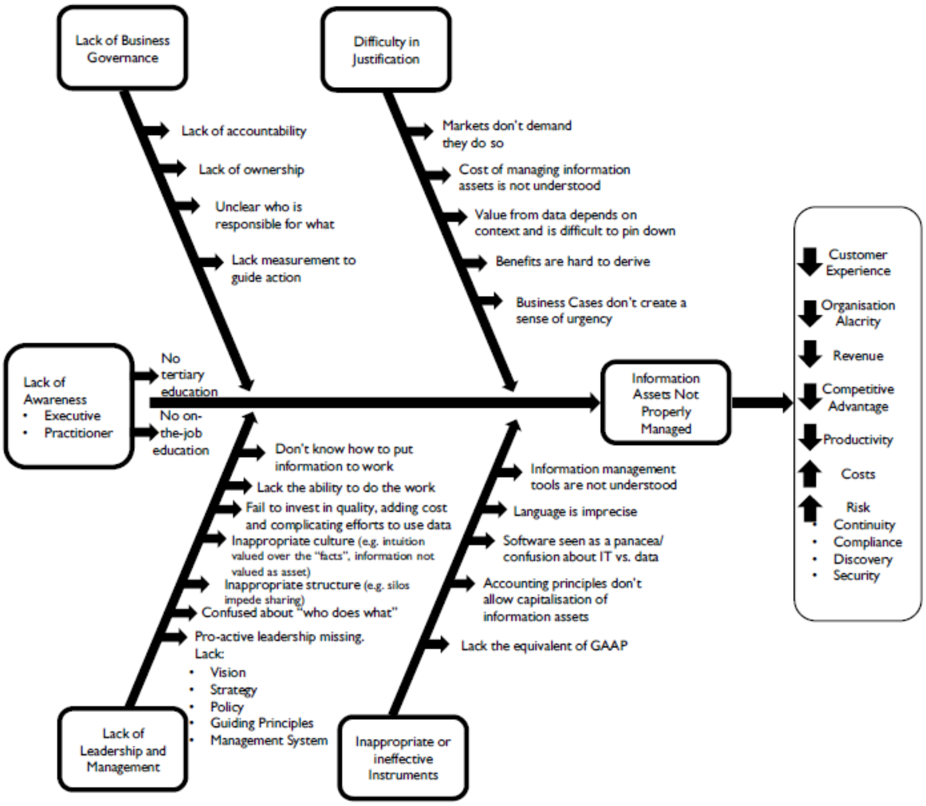
1.8 Common Causes of Data Quality Issues

Data quality issues can arise at any time in the lifecycle, and may have multiple causes: Data entry, data processing, system design and manual intervention in automated processes.

1.8.1 Issues caused by lack of leadership

Barriers to effective management of data quality:

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


Lack of Business Governance
Lack of accountability
Lack of ownership
Unclear who is responsible for what
Lack measurement to guide action
Difficulty in Justification
Markets don't demand they do so
Cost of managing information assets is not understood
Value from data depends on context and is difficult to pin down
Benefits are hard to derive
Business Cases don't create a sense of urgency
Lack of Awareness
• Executive
• Practitioner
No tertiary education
No on-the-job education
Information Assets Not Properly Managed
Customer Experience
Organisation Alacrity
Revenue
Competitive Advantage
Productivity
Costs
Risk
• Continuity
• Compliance
• Discovery
• Security
Don't know how to put information to work
Lack the ability to do the work
Fail to invest in quality, adding cost and complicating efforts to use data
Inappropriate culture (e.g. intuition valued over the "facts", information not valued as asset)
Inappropriate structure (e.g. silos impede sharing)
Confused about "who does what"
Pro-active leadership missing.
Lack:
• Vision
• Strategy
• Policy
• Guiding Principles
• Management System
Lack of Leadership and Management
Information management tools are not understood
Language is imprecise
Software seen as a panacea/ confusion about IT vs. data
Accounting principles don't allow capitalisation of information assets
Lack the equivalent of GAAP
Inappropriate or ineffective Instruments

© 2017 dataleaders.org
Used with permission

Barriers that slow/hinder/prevent companies from managing their information as a business asset
Most commonly observed root causes
Danette McGilvray / James Price / Tom Redman
October 2016

Work based on research by Dr. Nina Evans and James Price, see "Barriers to the Effective Deployment of Information Assets" at www.dataleaders.org

1.8.2 Issues caused by data entry process

- **Data entry interface issues**: Needs to have edits and controls
- **List entry placement**: Order of values in a dropdown list
- **Field overloading**: Reusing fields for different business purposes over time creates confusion
- **Training Issues**: Awareness of the impact of incorrect data. Incentivise for accuracy, not speed.
- **Changes to business processes**: New business rules and data quality requirements need to be incorporated into systems.
- **Inconsistent business process execution**: will produce inconsistent data

1.8.3 Issues caused by data processing functions

- **Incorrect assumptions about data sources**: Not enough known about data sources, details are missed
- **Stale business rules:** Business rules may change over time, and should be periodically reviewed
- **Changed data structures:** Source changes without informing downstream

1.8.4 Issues caused by system design

- **Failure to enforce referential integrity**:

  - Duplicate data that breaks uniqueness rules
  - Orphan rows, in some reports and not others, leading to different 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 result in overstated aggregation
- **Coding inaccuracies and gaps:** Data mapping or rules for processing incorrect
- **Data model inaccuracies:** assumptions in data model must be supported by actual data
- **Field overloading**
- **Temporal data mismatches:** Need a consolidated data dictionary
- **Weak Master Data Management:** Choose unreliable data sources
- **Data duplication:** Two main types:
  - **Single source – Multiple Local instances:** e.g. instances of the same customer in multiple tables
  - **Multiple sources – Single Instance:** Data instances with multiple authoritative sources

1.8.5 Issues caused by fixing issues

Manual data patches directly to database. Most change the data in place, and can only be undone by a database restore.

1.9 Data Profiling

Data profiling is a form of data analysis used to inspect the data and assess data quality. A profiling engine produces statistics that can be analysed to identify patterns in data content and structure.

- **Counts of NULLS:** Inspect to see if they are allowed
- **Max/Min value:** Identifies outliers
- **Max/Min Length:** Outliers in fields for specific length values
- **Frequency distribution** of values for individual columns
- **Data Type and Format:** Non-conformance or unexpected formats

1.10 Data Quality and Data Processing

1.10.1 Data Cleansing

Data cleansing or Scrubbing transforms data to conform to data standards or domain rules. Detecting and correcting errors. The need for data cleansing can be addressed by:

- Implementing controls to prevent data entry errors
- Correcting the data in the source system
- Improving the business processes that create data

1.10.2 Data enhancement

Data enhancement or enrichment is the process of adding attributes to data to increase its quality or usability.

- **Time/Date Stamp:**
- **Audit data:** adds lineage
- **Reference vocabularies**
- **Contextual information:**
- **Geographic information:** Address standardisation and geocoding

- **Demographic information:** customer data enhanced through demographic information
- **Psychographic information:** segment target populations by behaviours
- **Valuation information:** Assets

1.10.3 Data parsing and formatting

Data Parsing is the process of analysing data using pre-determined rules to define its content or value.

1.10.4 Data Transformation and Standardisation

Rule based transformations map data values in their original formats and patterns into target representation. Standardisation employs rules that capture context, linguistics and idioms recognised as common over time.

2 Activities

2.1 Define High Quality Data

High quality data is fit for the purposes of data consumers. Understand business needs, define terms, identify pain points and start to find consensus about drivers and data quality priorities. Ask questions of stakeholders to determine the understanding of the business benefits of high-quality data and the impact of poor quality data.

Understand the current state of data quality:

- Understand business strategy and goals
- Pain points, risks and business drivers
- Direct assessment of data (profiling)
- Documentation of data dependencies in business processes

Prioritise opportunities based on benefits to the organisation.

2.2 Define a Data Quality Strategy

Data quality priorities must align with business strategy. Develop a framework which includes methods to:

- Understand and prioritise business needs
- Identify data critical to business needs
- Define business rules and data quality standards based on business requirements
- Assess data against expectations
- Share findings and get feedback from shareholders
- Prioritise and manage issues
- Identify and prioritise opportunities for improvement
- Measure, monitor and report on data quality
- Manage Metadata produced through quality processes
- Integrate data quality controls into business and technical processes

2.3 Identify Critical Data and Business Rules

- Critical data
  - If it were higher quality would provide greater value
  - Regulatory requirements
  - Financial value

  - Direct impact to customers
  - Start with Master Data
  - Ranked list of data for DQ team
- Business Rules
  - Describe expectations about the quality of the data
  - Often not documented – need to reverse engineer
  - How data is collected or created
  - Measurements describe if data is fit for use
  - Discovery and refinement of rules is an ongoing process

2.4 Perform an Initial Data Quality Assessment

Query the data to understand content and relationships. Data stewards, SMEs, data consumers and DQ analysts prioritise findings.

The goal of the initial assessment is to learn about the data to make an actionable plan for improvement:

- Define the goals to drive the work
- Identify data to be assessed. Start small
- Identify uses of the data and consumers of the data
- Identify known risks of the data
- Inspect data based on known and proposed rules
- Document levels of non-conformance and types of issues
- Perform in depth analysis to prioritise issues and develop root cause hypotheses
- Confirm issues and priorities with stakeholders
- Planning

2.5 Identify and Prioritise Potential Improvements

- Prioritise actions based on business impact
- Develop preventative and corrective actions
- Confirm planned actions with stakeholders
- Large-scale profiling efforts should focus on the most critical data
- Profiling identifies issues, but not root causes

2.6 Define Goals for Data Quality Improvement

- Quick hits as well as long term strategic changes
- Address root causes
- Set achievable goals based on quantification of the business value of DQ improvements
- Determine ROI of fixes of issues based on:
  - Criticality of the data
  - Amount of affected data
  - Age of the data
  - Number and type of business processes impacted
  - No of stakeholders affected
  - Associated risks
  - Cost of root cause remediation
  - Cost of work arounds

2.7 Develop and Deploy Data Quality Operations

2.7.1 Manage Data Quality rules

Define rules up front to:

- Set clear expectations
- Provide requirements for edits and controls to prevent data issues from being introduced
- Provide DQ requirements to vendors
- Foundation for DQ measuring

DQ rules should be managed as Metadata and should be:

- **Documented consistently:** Templates
- **Defined in terms of DQ Dimensions:** Help people understand what is being measured
- **Tied to business impact:** Connect standards and rules to organisational success
- **Backed by data analysis:** Test rules on actual data
- **Confirmed by SMEs**
- **Accessible to all data consumers**

2.7.2 Measure and monitor Data Quality

Two reasons to implement operational data quality measurements:

- To inform consumers about levels of quality
- Manage risk that may be introduced by technical or business changes

Express as a percentage where (r) is the rule being tested.

$$ValidDQI(r) = \frac{\left(TestExecutions(r) - ExceptionsFound(r)\right)}{TestExecutions(r)}$$

$$InvalidDQI(r) = \frac{\left(ExceptionsFound(r)\right)}{TestExecutions(r)}$$

Example:

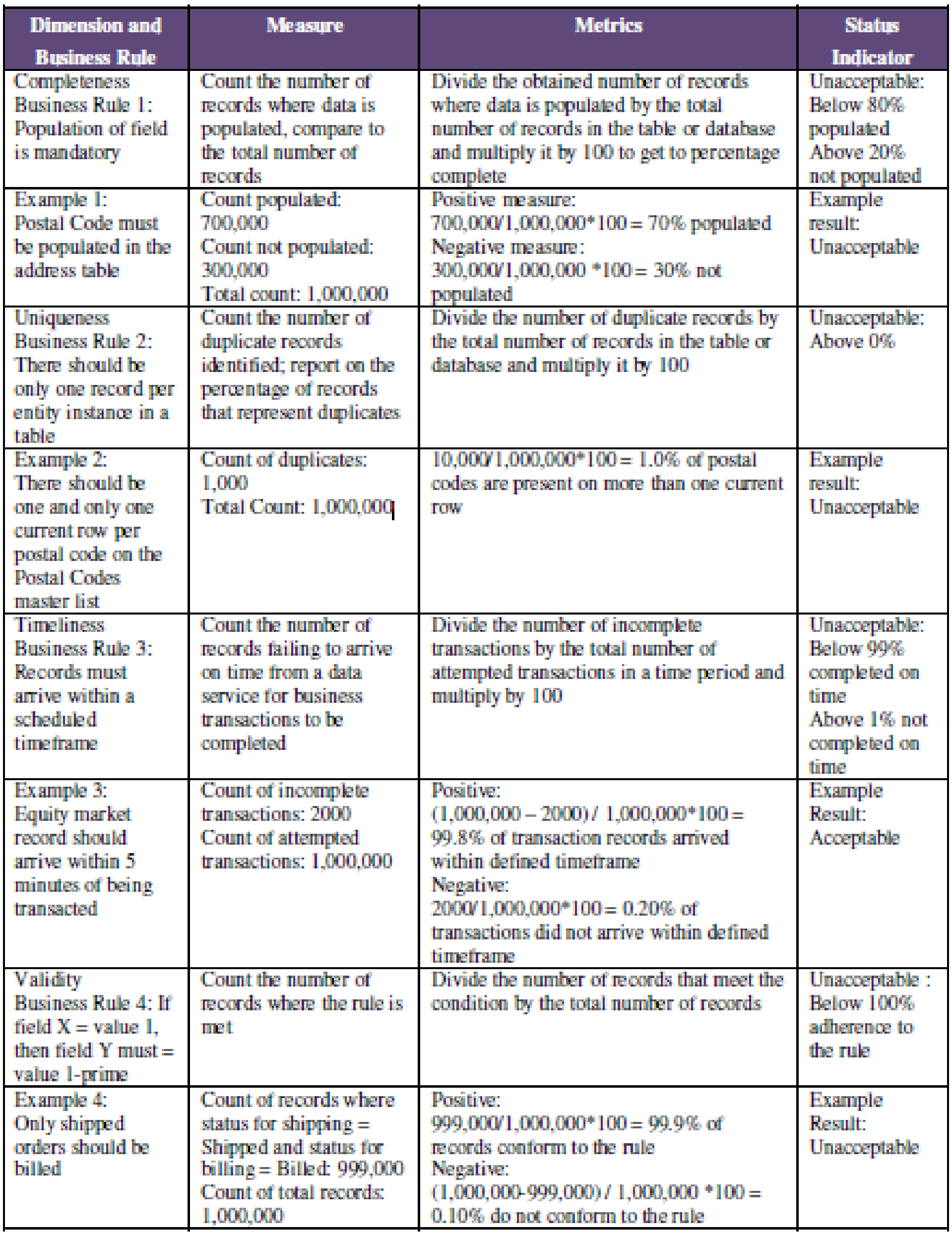
10000 tests of business rule (r) found 560 exceptions

Therefore: Valid DQ = (10000-560)/10000 = 94.4%

Invalid DQ = 560/10000 = 5.6%

Organise results in a table as shown below:

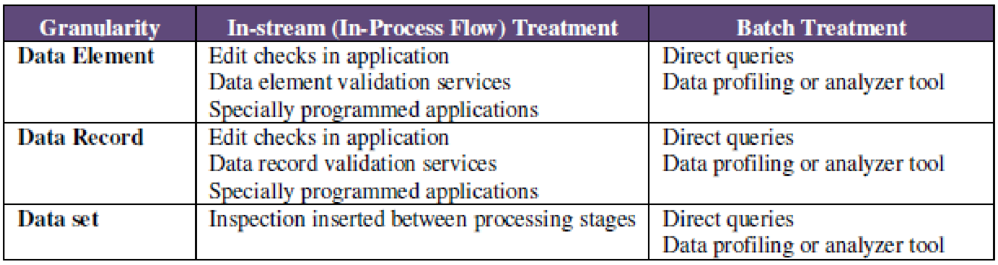
Table 30 DQ Metric Examples



Dimension and Business Rule	Measure	Metrics	Status Indicator
Completeness Business Rule 1: Population of field is mandatory	Count the number of records where data is populated, compare to the total number of records	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	Unacceptable: Below 80% populated Above 20% not populated
Example 1: Postal Code must be populated in the address table	Count populated: 700,000 Count not populated: 300,000 Total count: 1,000,000	Positive measure: 700,000/1,000,000*100 = 70% populated Negative measure: 300,000/1,000,000 *100 = 30% not populated	Example result: Unacceptable
Uniqueness Business Rule 2: There should be only one record per entity instance in a table	Count the number of duplicate records identified; report on the percentage of records that represent duplicates	Divide the number of duplicate records by the total number of records in the table or database and multiply it by 100	Unacceptable: Above 0%
Example 2: There should be one and only one current row per postal code on the Postal Codes master list	Count of duplicates: 1,000 Total Count: 1,000,000	10,000/1,000,000*100 = 1.0% of postal codes are present on more than one current row	Example result: Unacceptable
Timeliness Business Rule 3: Records must arrive within a scheduled timeframe	Count the number of records failing to arrive on time from a data service for business transactions to be completed	Divide the number of incomplete transactions by the total number of attempted transactions in a time period and multiply by 100	Unacceptable: Below 99% completed on time Above 1% not completed on time
Example 3: Equity market record should arrive within 5 minutes of being transacted	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000	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	Example Result: Acceptable
Validity Business Rule 4: If field X = value 1, then field Y must = value 1-prime	Count the number of records where the rule is met	Divide the number of records that meet the condition by the total number of records	Unacceptable : Below 100% adherence to the rule
Example 4: Only shipped orders should be billed	Count of records where status for shipping = Shipped and status for billing = Billed: 999,000 Count of total records: 1,000,000	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 = 0.10% do not conform to the rule	Example Result: Unacceptable

Measurements can be taken at three levels of granularity:

Table 31 Data Quality Monitoring Techniques



Granularity	In-stream (In-Process Flow) Treatment	Batch Treatment
Data Element	Edit checks in application Data element validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data Record	Edit checks in application Data record validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data set	Inspection inserted between processing stages	Direct queries Data profiling or analyzer tool

2.7.3 Develop Operational Procedures for Managing Data Issues

The DQ Team must respond to issues timeously and effectively. Design and implement operational procedures for:

- **Diagnosing issues:** Root cause analysis requires input from technical and business SMEs
  - Review issues in context of processing flows to isolate point where flaw is introduced
  - Evaluate whether there been environment changes that could have caused errors
  - Evaluate whether other process issues contributed
  - Determine whether there are issues with external data that could have affected this data
- **Formulating options for remediation:** Based on diagnosis, evaluate alternatives for addressing the issue:
  - Non-technical such as lack of training, leadership support, accountability
  - Modification of systems to eliminate root causes
  - Developing controls to prevent issue
  - Additional inspecting and monitoring
  - Directly correcting flawed data
  - Take no action based on cost and impact of correction versus the value of data correction.
- **Resolving issues:** DQ team and business data owners determine the best way to solve issues

An incident tracking system should be used:

- Standardise data quality issues and activities
- Provide an assignment process for data issues
- Manage issue escalation procedures
- Manage data quality resolution workflow

2.7.4 Establish Data Quality Service Level Agreements

A DQ SLA specifies the organisation's expectation for response and remediation for DQ issues in each system. The SLA defines roles and responsibilities associated with performance DQ procedures.

SLA establishes time limits for notification generation, the names in the management chain and when escalation should occur.

SLA reporting can be scheduled or driven by business.

2.7.5 Develop Data Quality Reporting

Reporting should focus around:

- DQ Scorecard
- DQ trends
- SLA metrics
- DQ Issue management focussing on the status of issues and resolutions
- Conformance of the DQ team to governance policies
- Conformance of IT and business teams to DQ policies
- Positive effects of improvement projects

3 Tools

- **Data Profiling Tools:**
  - Produce high level statistics to identify patterns in the data
  - Perform initial assessment of DQ

  - Enable assessment of large data sets
- **Data Querying Tools:**
  - Query more deeply to answer questions raised by profiling
- **Modelling and ETL Tools:**
  - Can be detrimental to data if used without knowledge of the data
  - Can improve quality if used with the data in mind
- **Data Quality Rule Templates**
- **Metadata Repositories:** Definitions of high quality data is Metadata

4 Techniques

4.1 Preventative Actions

Ways to prevent poor quality data entering an organisation:

- **Establish data entry controls:** Data entry rules
- **Train data producers:** Value accuracy rather than speed
- **Define and enforce rules:** Create a 'data firewall', a table with all the business rules, to check quality before the data is used
- **Demand high quality data from data suppliers:** Examine processes to check structures, data sources and provenance
- **Implement data governance and stewardship:**
- **Institute formal change control:** Ensure changes are tested before implementing

4.2 Corrective Actions

Perform data correction in three general ways:

- **Automated correction:**
  - Rule based standardisation, normalisation and correction
  - No manual intervention
  - Requires environment with well-defined standards, rules and known error patterns.
- **Manually-driven correction:**
  - Use automated tools with manual review before committing modified values to persistent storage.
  - Environments where data sets require human oversight e.g. MDM
- **Manual correction:**
  - Best done through an interface with controls and edits which produces an audit trail.
  - Avoid making manual corrections directly to the production environment

4.3 Quality Check and Audit Code Modules

4.4 Effective Data Quality Metrics

Characteristics of informative metrics:

- Measurability
- Quantifiable within a discrete range
- Business relevance
- Correlate with the influence of the data on the key business expectations
- Acceptability
- Data meets business expectation based on acceptability thresholds
- Accountability/Stewardship

- Understood and approved by key stakeholders
- Controllability
- Should trigger an action to improve data
- Trending
- Measure DQ over time

4.5 Statistical Process Control

SPC is a method to manage processes by analysing measurements of variation in process inputs, outputs or steps.

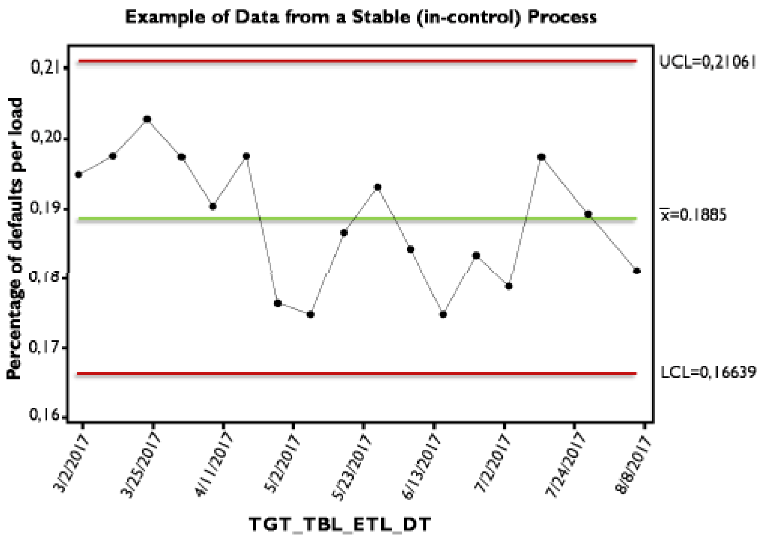
A process is a series of steps to turn inputs to outputs. SPC is based on the assumption that when a process with consistent inputs is executed consistently it will produce consistent outputs.

Uses measures of central tendency (mean, median, mode) and variability around a central value (range, variance, standard deviation) to establish tolerances for variances within a process

Two types of variance:

- Common causes:
  - Inherent in the process
  - Process is in statistical control when only common causes, and range of variation (baseline) is established.
- Special causes: Unpredictable or intermittent

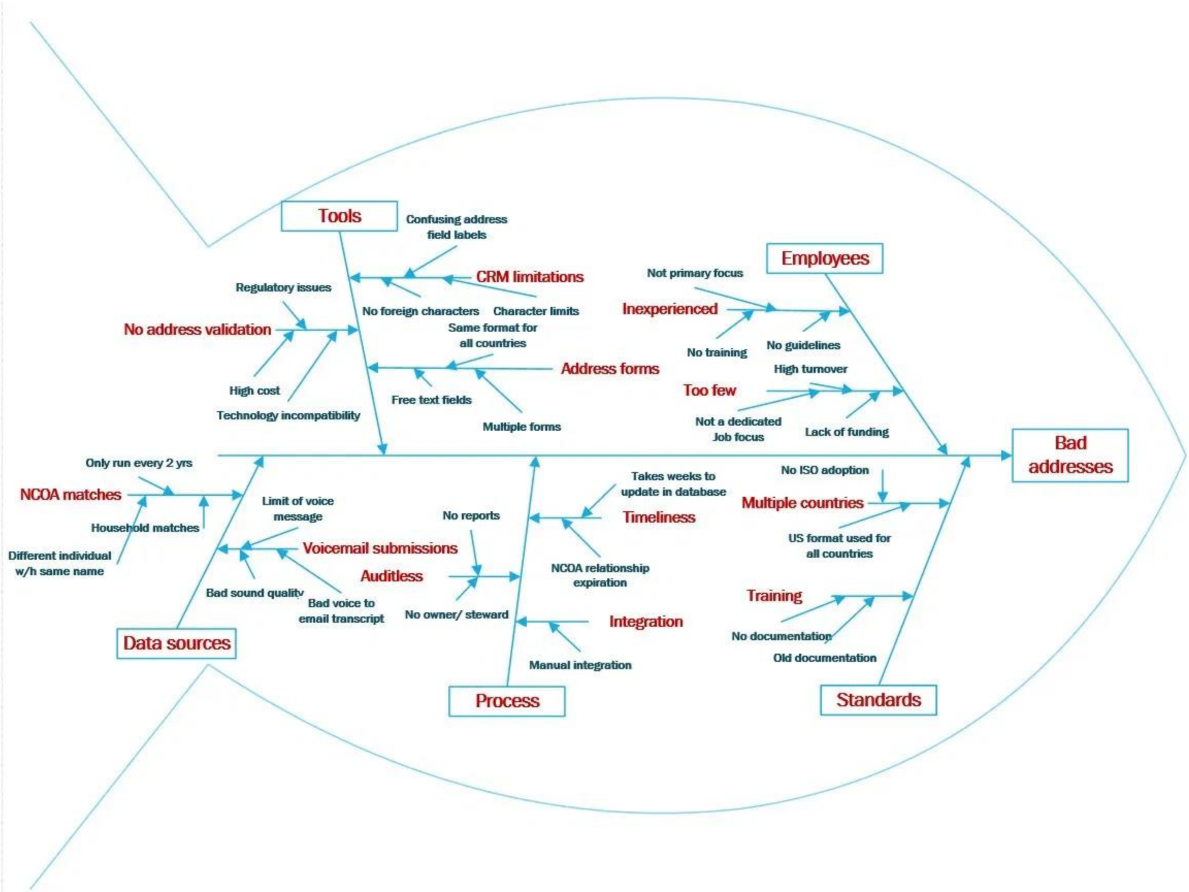
SPC is used for control, detection and improvement. Early detection of unexpected variation simplifies root cause investigation.



4.6 Root cause Analysis

Common techniques:

- Pareto analysis (the 80/20 rule)
- Fishbone diagram analysis
- Track and trace
- Process analysis
- Five Whys (McGilvray, 2008)



The Ishikawa / Fish Bone Diagram

5 Implementation Guidelines

Improving DQ in an organisation requires changing how the people think and behave towards data. A DQ implementation needs to plan for:

- **Metrics on the value of data and the cost of poor data:** To raise organisational awareness and funding for improvements
- **Operating mode for IT/Business interactions:** Business people know how important the data is and IT can translate definitions of data quality into queries that identify records which don't comply
- **Changes in how projects are executed:** Identify issues early and build data quality expectations into projects
- **Changes to business processes:** DQ team assesses processes and recommends changes
- **Funding for remediation and improvement projects:** Document costs and benefits of fixing data so that it can be prioritised
- **Funding for DQ operations:** Ongoing monitoring, reporting and fixing DQ issues

5.1 Readiness Assessment/Risk Assessment

Consider the following to assess the organisational readiness to accept DQ practices:

- **Management commitment to managing data as a strategic asset:**
  - How much do they know about data as an asset, risks of poor-quality data, importance of data governance?
- **The organisation's current understanding of the quality of its data:**

    - Pain points – helps identify and prioritise improvement projects
- **The actual state of the data:**
    - Profiling, analysis and quantification of known pain points
- **Risks associated with data creation, processing or use:**
    - Identify what can go wrong, and the potential damage to the organisation
- **Cultural and technical readiness for scalable data quality monitoring:**
    - Requires a good collaborative relationship between IT and business teams

5.2 Organisation and Cultural Change

Promote awareness of the role and importance of data in the organisation.

All employees raise DQ issues, ask for good quality data as consumers, and provide good quality data to others. Every person who touches the data can impact its quality.

Employees need to think and act differently to produce and manage better quality data. This requires training focussing on:

- Common causes of data problems
- Relationships and why DQ requires an enterprise approach
- Consequences of poor quality data
- Necessity for ongoing improvement
- Becoming data lingual
- Introduce process changes

6 Data Quality and Data Governance

A data quality program is more effective when part of a data governance program.

6.1 Data Quality Policy

All Data Management Knowledge Areas require a data policy, especially if they touch on regulatory areas:

- Purpose, scope and applicability of the policy
- Definition of terms
- Responsibilities of the Data Quality program
- Responsibilities of other stakeholders
- Reporting
- Implementation of the policy, including links to risk, preventative measures, compliance, data protection and data security.

6.2 Metrics

High level categories of DQ metrics:

- **Return on investment:**
- **Levels of quality:**
- **Data Quality trends:**
- **Data issue management metrics:**
- **Conformance to service levels:**
- **Data Quality plan rollout:**