Data Management

Data Quality

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL
UNIVERSITY DUBLIN

Dr Emma Murphy

Week 8, March 2020

Recap

- Ethical Theories
- Data Lifecycle
- Data Management Strategies
- Data Governance
- Data Privacy

Topics

- Ethical Theories
- Data Lifecycle
- Data Management Strategies
- Data Governance
- Data Privacy
- Data Quality
- Data Bias
- Data Security

Data Quality

- The term *data quality* refers both to the characteristics associated with high quality data and to the processes used to measure or improve the quality of data.

- (DAMA, 2017)

Defining DQ

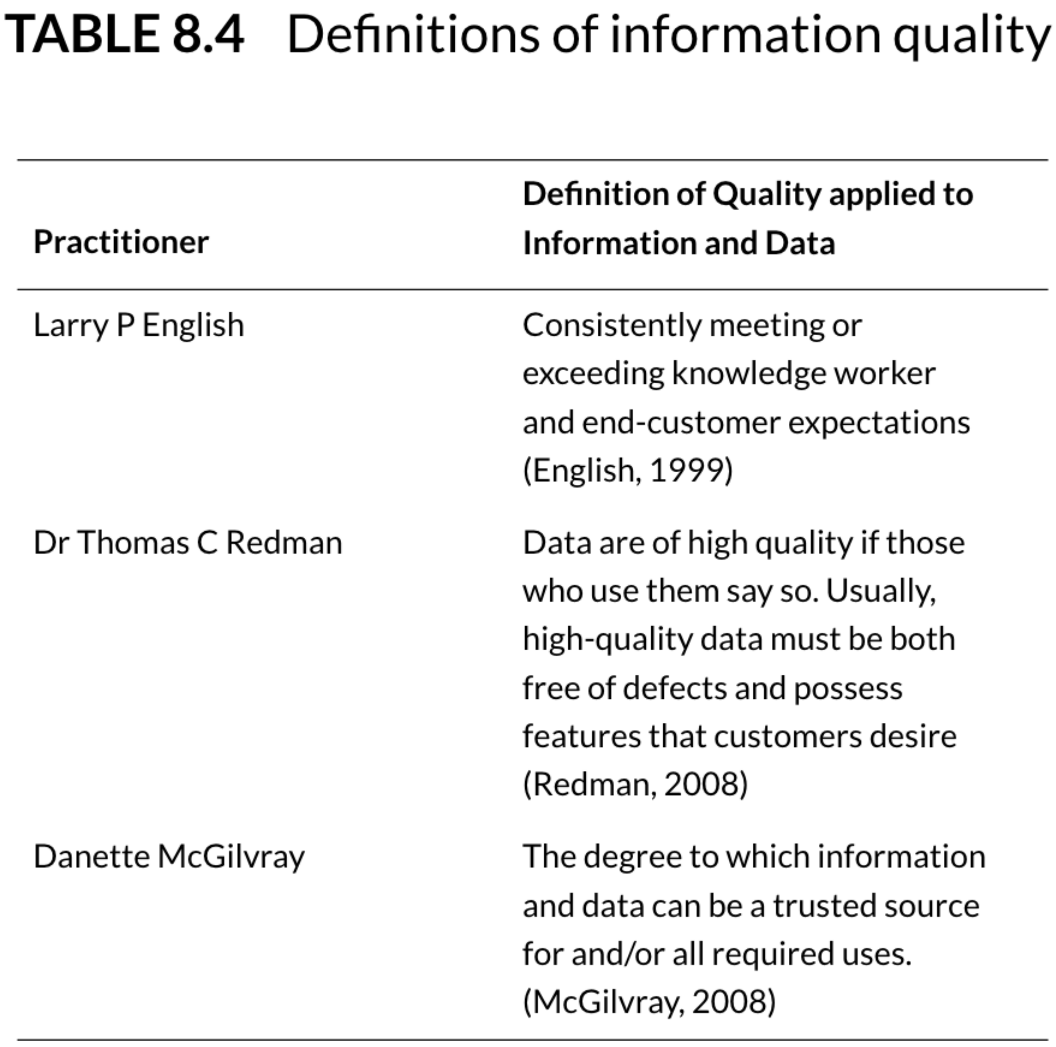


TABLE 8.4 Definitions of information quality

Practitioner	Definition of Quality applied to Information and Data
Larry P English	Consistently meeting or exceeding knowledge worker and end-customer expectations (English, 1999)
Dr Thomas C Redman	Data are of high quality if those who use them say so. Usually, high-quality data must be both free of defects and possess features that customers desire (Redman, 2008)
Danette McGilvray	The degree to which information and data can be a trusted source for and/or all required uses. (McGilvray, 2008)

Data Quality Management

- Just as with ethics, information quality management requires you to have some mechanism for defining and measuring what ‘good’ is.
- It also requires formal processes and methods to plan for, design for, and ensure controls for good-quality information.

(O’Keefe and O’Brien, 2018)

Context

- Data is of high quality to the degree that it meets the expectations and needs of data consumers. That is, if the data is fit for the purposes to which they want to apply it. It is of low quality if it is not fit for those purposes. Data quality is thus dependent on context and on the needs of the data consumer.

(DAMA, 2017)

Challenges

- One of the challenges in managing the quality of data is that expectations related to quality are not always known.
- However, if data is to be reliable and trustworthy, then data management professionals need to better understand their customers' quality requirements and how to measure them.

(DAMA, 2017)

Critical Data

- Most organizations have a lot of data, not all of which is of equal importance.
- One 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 program scope and focus and enables it to make a direct, measurable impact on business needs.

(DAMA, 2017)

Defining Critical Data Elements

- Critical data elements are those on which the success of business processes and corresponding business applications rely. Yet of the thousands of data elements that could exist within an organization, how would one distinguish critical data elements from your everyday, run-of-the-mill data elements?

https://www.sciencedirect.com/topics/computer-science/critical-data-element

Defining Critical Data Elements

- These need to be defined within the organization and context. For example:
  - “containing personal information protected under a defined privacy or confidentiality policy”
  - “containing critical information about an employee”
  - “containing critical information about a supplier”
  - “containing detailed information about a product”
  - “required for operational decision processing”

https://www.sciencedirect.com/topics/computer-science/critical-data-element

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 results of initial data quality assessment as well as ongoing measurement.

Data Quality Dimensions

- In order to measure the quality of data, an organization needs to establish characteristics 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.

DAMA, 2017

Data Quality Dimensions

- For example, if the data in the customer email address field is incomplete, then we will not be able to send product information to our customers via email, and we will lose potential sales.
- Therefore, we will measure the percentage of customers for whom we have usable email addresses, and we will improve our processes until we have a usable email address for at least 98% of our customers.

(DAMA, 2017)

Data Quality Dimensions

- Dimensions include some characteristics that 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.

DAMA Data Quality Dimensions

- While there is not a single, agreed-to set of data quality dimensions, DAMA (2017) has created a set of dimensions based on common ideas and describes approaches to measuring them.

(DAMA, 2017)

Dimension 1 - Accuracy

- Accuracy refers to the degree that data correctly represents “real life” entities.

- Accuracy can be difficult to measure, unless an organisation can reproduce data collection or manually confirm accuracy of records.

(DAMA, 2017)

Dimension 2: Completeness

- Completeness refers to whether all required data is present. Completeness can be measured at the data set, record, or column level.
- 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 levels may require comparison to a source of record or may be based on historical levels of population.

Dimension 3: 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.

Dimension 3: Consistency

- Consistency may be defined between one set of attribute values and another attribute set within the same record (record-level consistency), 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).

- Characteristics that are expected to be consistent within and across data sets can be used as the basis for standardising data. Data standardisation refers to the conditioning of input data to ensure that data meets rules for content and format.

Dimension 4: Data 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.

Dimension 5: 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

Dimension 6: 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.

(DAMA, 2017)

Dimension 7: Uniqueness

- 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.

Dimension 8: 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.
- 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.

Key quality dimensions

- Organizations need to identify what dimensions are important to business processes and can also be measured.
- The key aspect of what is being measured is that it should be linked back to a business risk or issue.
  - For example, if the blood group is not recorded correctly for patients in a hospital, this could result in potential treatment complications. Therefore, a hospital might require 100 per cent population of blood group and an accuracy level of 99.999 per cent for that data.

Question?

- What are the key quality dimensions that a clinician would be concerned with if asked to base a diagnosis or care plan for a patient based on commercial personal sensing devices such as the fitbit or Apple health watch?
  - Using data such as hours of sleep, heart rate, step count.

- *[Accuracy/ Completeness/ Consistency/ Data Integrity/ Reasonability/ Timeliness/ Uniqueness/ Validity]*

Business Rules

- Data quality dimensions are used as inputs into data quality business rules. These rules describe how the data should be for it to be useful.
- They codify the expectation of quality of data and act either to prevent, detect or remedy issues when they are identified.

Data Profiling

- Data profiling is the process by which data is measured against defined business rules to identify issues in the data.
- A profiling tool will produce a statistical analysis of the data, which can be used to inform analysts about patterns of defect in the content and structure of the data. Depending on the business rule a value may be significant or not.

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. A profiling engine can produce statistics that analysts can use to identify patterns in data content and structure. For example:
  - Counts of Nulls
  - Max/Min Value
  - Max/Min Length
  - Frequency Distribution
  - Data Type and Format
  - Profiling also includes cross-column analysis and inter-table analysis.

Data Profiling

- For example, if a data-profiling exercise is expected to reveal that 100 per cent of all the rows in field X are null, but there are 20 per cent of the records that have values, this might indicate a problem in the data that needs to be investigated.

Data Cleansing

- 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 through cleansing. Ideally, the need for data cleansing should decrease over time as the root causes of data quality issues are resolved. 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

Meta Data

- Metadata is critical to managing the quality of data. The quality of data is based on how well it meets the requirements of data consumers. 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.

Meta data

- Well-managed Metadata can also support the effort to improve the quality of data. 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.

Data Quality Improvement

- Most approaches to improving data quality are based on the techniques of quality improvement in the management of physical products. In this paradigm, data is understood as the product of a set of processes. At it's simplest, a process is defined as a set of steps that turn inputs into outputs.

Data Quality Improvement

- 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.

Data Quality Improvement

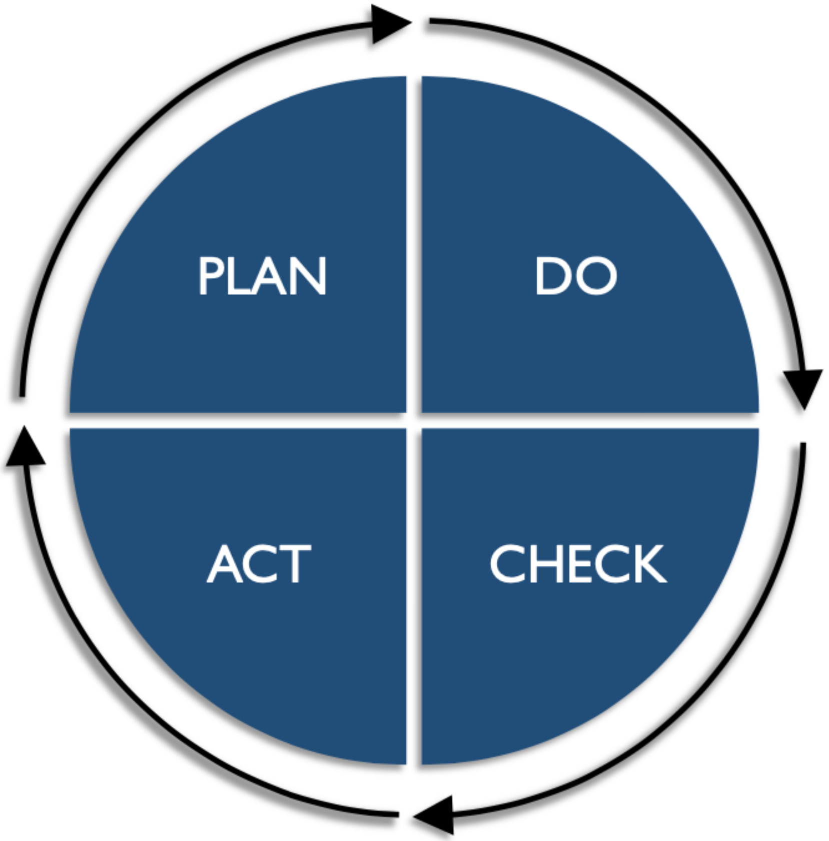
- For a given data quality set, the Data Quality lifecycle begins by identifying data that does not meet data consumer's requirements and data issues that are obstacles to the achievement of business objectives.
- Data needs to be assessed against key dimensions of quality and known business requirements. 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 Data Stewards and other stakeholders.

PDCA Model

The PDCA Model is a four-stage model for continuous improvement that can be used in areas including:

- Business Process Management
- Product Lifecycle Management
- Project Management
- Human Resource Management

The model was developed by Dr Edward Deming, considered the father of modern quality control, and is also known as the Deming Model.



(DAMA, 2017)

Plan

- 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.

(DAMA, 2017)

Do

- In the *Do* stage, the DQ 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 DQ team can work with process owners to implement changes. For root causes that require technical changes, the DQ team should work with technical teams to ensure that requirements are implemented correctly and that technical changes do not introduce errors.

(DAMA, 2017)

Check

- The *Check* stage involves actively monitoring the quality of data as measured against requirements. As long as data meets defined thresholds for quality, additional actions 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.

(DAMA, 2017)

Act

- The *Act* stage is for activities to address and resolve emerging data quality issues. The cycle restarts, as the causes of issues are assessed and solutions proposed. 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

(DAMA, 2017)

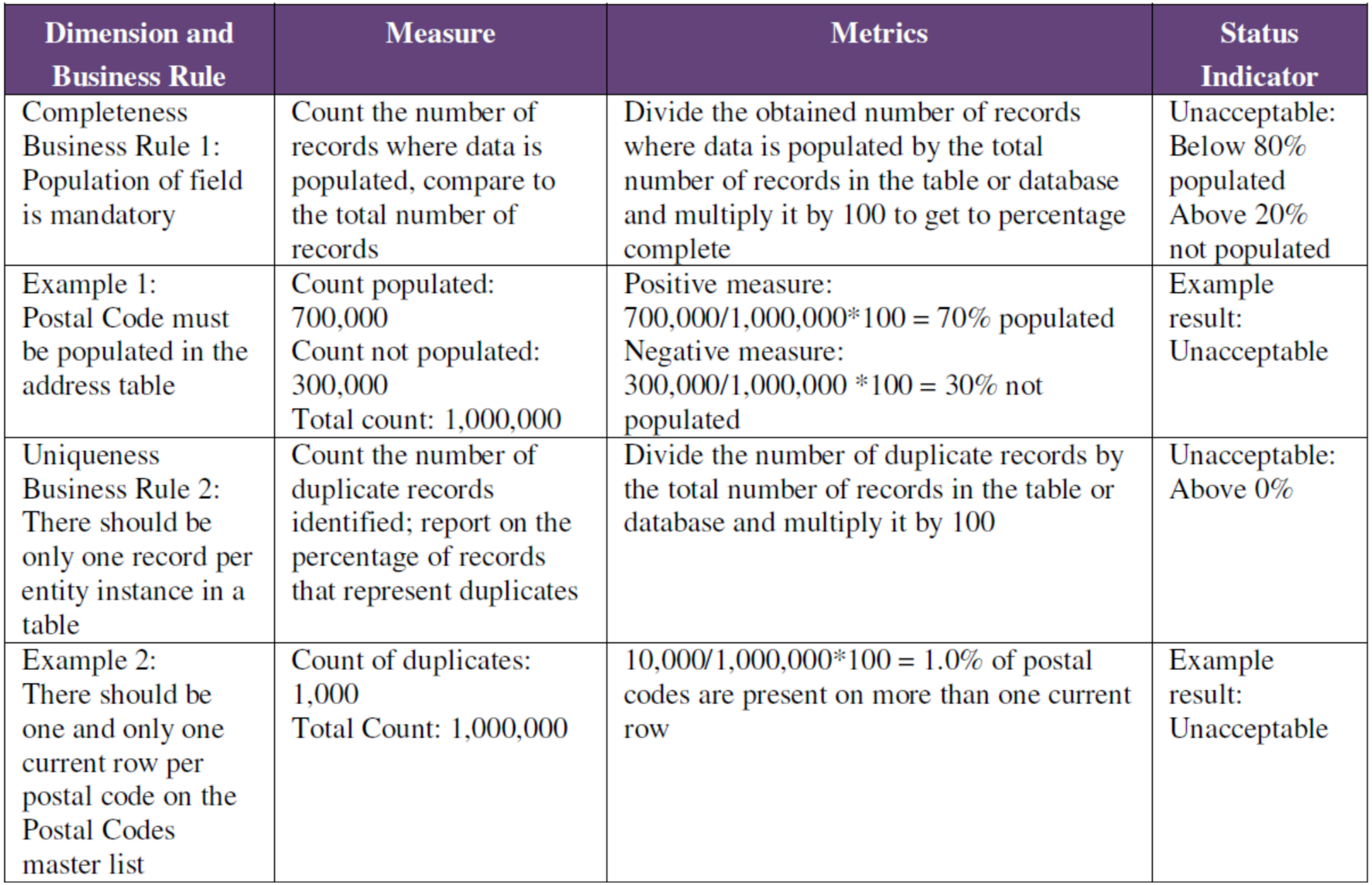
Measuring Data Quality

- The quality of information is something that can be, to a greater or lesser extent, objectively measured.
- The importance of having a metric of quality is that it allows you to demonstrate the scale of a problem, and to demonstrate improvement arising from remedial actions you might take to improve the quality of information in your organization.
- (O'Keefe and O'Brien, 2018)

Data Quality Measurement

- The operational data quality management procedures depend on the ability to measure and monitor data quality. There are two primary reasons to implement operational measurements:
    - To inform consumers about levels of quality
    - To manage risks that change may be introduced
- Measurements should be developed based on findings from data assessment and root cause analysis. Knowledge of past problems should be applied to manage risk.
- Measurement rules can be described at two levels, the detailed related to the execution of individual rules and the overall, based on the aggregate.

Measurement 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

DQ Measurement Example

Dimension	Measure	Metric	Status Indicator
Accuracy: Heart rate sensor	Check data point within (clinically) possible data range	Work out maximum range including resting and active heartrate (i.e. 20-300bpm)	Acceptable if between range identified otherwise unacceptable

What are the risks when creating measures such as the one above?

How could they be overcome?

Causes of poor DQ

- A lack of leadership and appropriate governance;
- Difficulty in justifying the required improvements; inappropriate or ineffective tools to measure the value of information; data-entry process issues, including inconsistent processes and lack of training; data-processing issues, including assumptions about data sources; data systems design issues, including data modelling issues; issues caused by hasty fixes.

Cost

- The cost of getting data right the first time is cheaper than the costs from getting data wrong and fixing it later.
- Building quality into the data management processes from the beginning costs less than retrofitting it.
- Maintaining high quality data throughout the data lifecycle is less risky than trying to improve quality in an existing process. It also creates a far lower impact on the organization.
- Establishing criteria for data quality at the beginning of a process or system build is one sign of a mature Data Management Organization. Doing so takes governance and discipline, as well as cross-functional collaboration.

An Ethical Approach to DQ

- Organizations that are thinking in terms of quality systems and quality approaches tend to:
- Focus on the needs of their customers – they make sure they identify and understand them, both customers and their needs, correctly. Drive out fear and encourage pride in a job well done – blaming people for errors and failures of the overall system is counterproductive. While it might be tempting to blame the person nearest the symptom, often the root cause lies elsewhere in the organization. Focus on continuous improvement – by incrementally improving quality and not resting on laurels, the organization can develop a sustainable level of quality.

An Ethical Approach to DQ

- Avoid creating the ‘hero ethic’ where staff are thrown at problems to fix them manually i.e. individual employees cleansing or reworking data as it is encountered
- A much better approach is to apply a quality ethic in the organization and focus on preventing the defects rather than just hurrying to scrap and rework data. This requires an ethical focus on the needs, wants and concerns of the stakeholders in that data, which could include internal and external customers, so that quality is designed in.

(O’Keefe and O’Brien, 2018)

An Ethical Approach to DQ

- Measuring the successful outcomes and not tracking people on the defects that have been created but on the number of records that meet the required standard.
- This changes the ethic of quality improvement away from punishment for deviation from standard towards one of reward and recognition for good-quality outputs.
- (O’Keefe and O’Brien, 2018)

An Ethical Approach to DQ

- This ties into the wider ethical concepts of preserving and promoting human dignity.
- If you constantly highlight the failings of the organization, teams within the organization, or single out individuals as being below par, you can impact on morale and feelings of dignity at the micro and macro level

(O'Keefe and O'Brien, 2018)

References

DAMA International (2017) DAMA DMBOK, DAMA DMBOK – Data Management Body of Knowledge, Technics Publications, New Jersey, pp 381–85

O'Keefe, K. and O'Brien, D. (2018) Ethical Data and Information Management: Concepts, Tools and Methods (1st. ed.). Kogan Page Ltd., GBR.

Hasselbach, G and Tranberg, P (2017) Data Ethics: The new competitive advantage, PubliShare, Copenhagen.