|  |
| --- |
| Executive Programme in Tax and Digital TransformationAsian Development Bank, World Bank Group and VIA University College  Data Quality Study Guide |



Table of Contents

[Module 1: Introduction to data quality 4](#_Toc82703219)

[Other resources 4](#_Toc82703220)

[Learning outcomes 4](#_Toc82703221)

[Key concepts 4](#_Toc82703222)

[What is data quality? 5](#_Toc82703223)

[Fitness for purpose 5](#_Toc82703224)

[The data triangle 6](#_Toc82703225)

[The circle of data quality 7](#_Toc82703226)

[Data quality principles 7](#_Toc82703227)

[Data quality in a data driven organisation 8](#_Toc82703228)

[Data quality considerations 9](#_Toc82703229)

[What is data quality management? 11](#_Toc82703230)

[Data ownership 13](#_Toc82703231)

[Data quality management features and benefits 13](#_Toc82703232)

[Data architecture 14](#_Toc82703233)

[MODULE 1 ACTIVITY 1-1 18](#_Toc82703234)

[Conclusion 18](#_Toc82703235)

[Module 2: Procedures, process and practices in data quality 19](#_Toc82703236)

[Other resources 19](#_Toc82703237)

[Learning outcomes 19](#_Toc82703238)

[Key concepts 19](#_Toc82703239)

[Introduction 20](#_Toc82703240)

[Data quality framework 20](#_Toc82703241)

[Data quality dimensions 22](#_Toc82703242)

[Dimension indicators and business rules 23](#_Toc82703243)

[Data quality dimension weighting 27](#_Toc82703244)

[Data quality grades 28](#_Toc82703245)

[Data quality standards 29](#_Toc82703246)

[Data quality matrix 30](#_Toc82703247)

[Data quality tools 32](#_Toc82703248)

[Conclusion 33](#_Toc82703249)

[Module 3: Benefits, assessments and challenges 34](#_Toc82703250)

[Other resources 34](#_Toc82703251)

[Learning outcomes 34](#_Toc82703252)

[Key concepts 34](#_Toc82703253)

[Introduction 35](#_Toc82703254)

[Data quality recap 35](#_Toc82703255)

[Benefits of data quality management 35](#_Toc82703256)

[Organisational Efficiencies 36](#_Toc82703257)

[Risk management 36](#_Toc82703258)

[Explore new opportunities 37](#_Toc82703259)

[Preparing for data quality 37](#_Toc82703260)

[Enterprise maturity assessments 37](#_Toc82703261)

[Challenges for data quality management 45](#_Toc82703262)

[Data quality and data ethics 45](#_Toc82703263)

[Data quality and data value 46](#_Toc82703264)

[Data quality issues of the future 47](#_Toc82703265)

[MODULE 3 ACTIVITY 3–1 48](#_Toc82703266)

[Conclusion 48](#_Toc82703267)

[GLOSSARY 49](#_Toc82703268)

# Module 1: Introduction to data quality

## Other resources

You will be referred to the OECD website at <http://www.oecd.org/>, and in particular, the OECD handbook for internationally comparative education statistics.

The following additional resources were also used

* Data Management book of Knowledge (DAMA) - [DMBoK - Data Management Body of Knowledge (dama.org)](https://www.dama.org/cpages/body-of-knowledge)
* Keith Gordon’s book, Principles of Data Management (2013)
* ISO 8000-61 the international standard that specifies a process reference mode for data quality management

## Learning outcomes

On completion of this module you should be able to:

* Understand the terms data quality and data quality management;
* Understand the complex data eco system;
* Understand the circle of data quality and the data quality triangle concepts
* Outline generally data quality principles that enable good data quality practice;
* Explain broadly the benefits of implementing a data quality management program.

## Key concepts

* data quality
* fit for purpose
* data quality management
* data ownership
* data lifecycle

Introduction

This module is an introduction to data quality as a concept and explains why effective data quality management is critical to embed within an organisation.

This module includes:

* Introduction to data quality and data quality management - what is it, why is it important and how can it assist your organisation.
* Sounds principles for data quality management and data ownership for responsibility and accountability for data quality.
* The data lifecycle and determining when you should conduct data quality activities

On completion of the module there will be an activity for participants to complete.

This course should be seen as a broad introduction to data quality.

# What is data quality?

The fundamental effect of data quality is the right data being available at the right time, for the right person, to the right user, to make a decision and achieve the right outcome.

Consider the statement that the “weather is bad”, what does this really mean? Does it mean it is raining, is it windy, is it cold, is it hot, is it humid? If the initial statement is used as a guide of quality then how does this statement provide insights to decisions, like building a structure.

So it follows that if someone states that they have poor quality data, that can be difficult to interpret without a better way of describing the nature of the data.

The lead on understanding the statement is the need to understand the purpose, and then fitting the data as it aligns to the purpose.

Data quality refers to the state of data as it relates to it’s intended purpose. Generally, data will be considered high quality when it is fit for purpose. Data quality management goal is to improve the quality of data to make sure it is fit for purpose.

## Fitness for purpose

In data quality management the term quality is an assessment of whether an item or activity conforms with the requirements and as such the purpose.

An example that illustrates this might be the choosing a window when building a house. One has to consider not just the window, but it’s colour, dimensions so it fits, style, and practicalities, all dimensions that would be natural in this situation.

To determine the quality of data it must be measured. Generally, data quality is measured by considering the dimensions of the data as they relate to the purpose of the data.

ISO 8000-8 builds a fundamental computer science to create a definitive framework for the characteristics and requirements of data. The framework identifies three types of data quality as being:

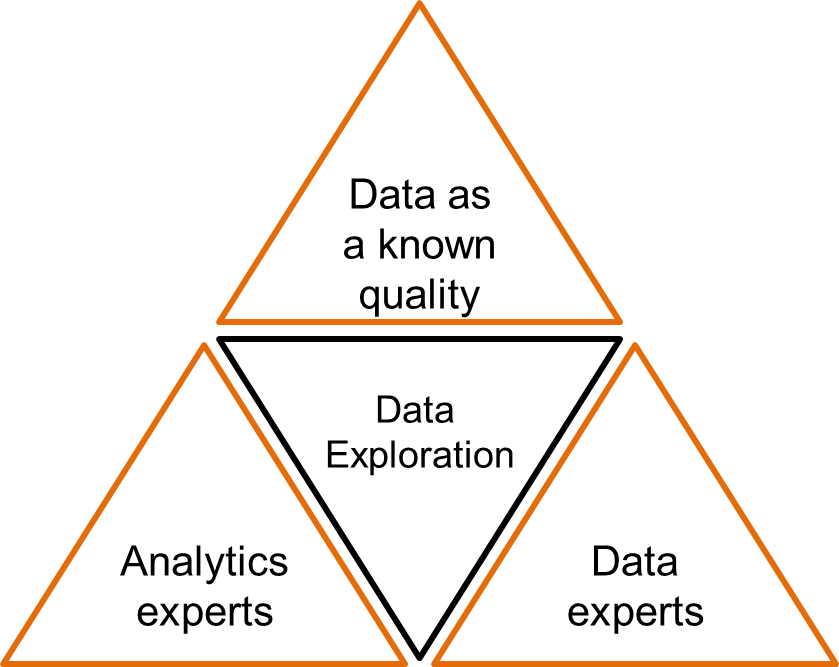
1. Syntactic (correct format for data)
2. Semantic (consistent common interpretation of the data)
3. Pragmatic (the data will be useful for the intended purpose of the user)

These three types can be abstract, so many people choose the more popular approach of data quality dimensions as detailed in the DAMA approach of data quality dimensions.

Data quality dimensions are covered in Module 3 of this study guide.

## The data triangle

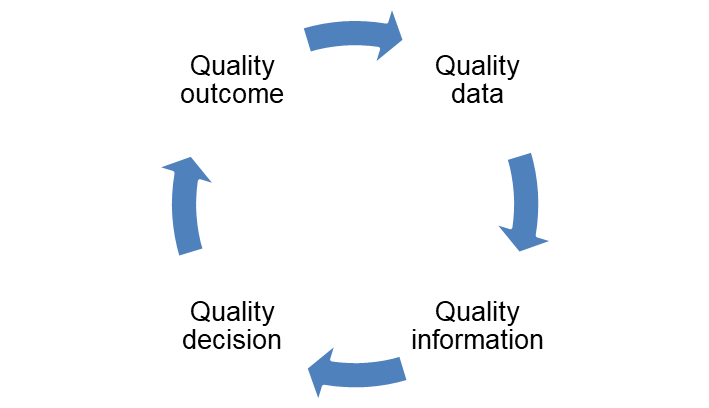
One unique way to look at the importance of data quality is a simple data triangle as illustrated below:



Each element can be described as follows:

* Top triangle – **data as a known quality** - it is important to have an understanding of the known quality of something before you commence the process of exploration.
* Middle triangle – **data exploration** – or exploration to insights helps better utilise good data for good decision making.
* Bottom left and right triangles – **analytics and data experts** - are functions that enable the data exploration to be valuable.

## The circle of data quality



As a general rule the quality data informs information that assists in better decisions that then lead to better outcomes, and may also include managing better data.

# Data quality principles

#### Consistent data quality management across the data lifecycle within business processes.

*demonstrating this principle:*

data quality process, detailed procedures and data quality tools and templates will support consistent data quality management across the organisation.

business processes will include data quality at various stages in the data lifecycle depending on the data source.

#### Data quality is measurable and assessed for fitness for purpose before use.

*demonstrating this principle:*

data quality dimensions and measurement criteria will support sound data quality metrics and when applied to the data quality standards will determine if the data is fit for the intended purpose.

#### Data quality management will consider prevention of data errors in the first instance by considering the costs and benefits of root-cause remediation.

*demonstrating this principle:*

the assessment stage of the data quality process involves analysis to identify the root cause of data quality issues. Preventing quality issues in the first instance is generally more cost effective than continual error correction activities. Cost benefit analysis is required prior to selecting and resourcing data quality improvements.

#### Data quality management will be objective and transparent, with data quality outcomes captured and made known to all data users.

*demonstrating this principle:*

data quality metrics captured in metadata which is accessible by all data users in the organisation. Data quality management will be governed independently.

#### Data quality is managed by dedicated roles that are responsible and accountable.

*demonstrating this principle:*

organisations data stewardship dictates who is responsible and accountable for data quality management.

# Data quality in a data driven organisation

Data has become a rich source of what is often described as the “new oil”. But what has been misunderstood is that not all oil is good oil and as such the need to collect ‘**all data**’ needs to be altered to collecting the ‘**right data’**.

It is also misunderstood that right data sits in isolation to the complexity of running a leading practice in being a data driven organisation which is again another misunderstood culture. Right data collection requires the input and interactions across a set number of strategic pillars, which include culture/people, law, process, technology and of course data.

Each of these aspects will be explained further down and will illustrate their interactions and their importance about being a data driven organisation.

Many people have contributed to a better understanding of a data driven business, but few understand the real challenges associated with shifting the theoretical to a “practice to action”, and it can be said:

*“The journey of digital and being a data driven organisation is a marathon without a finish line”.*

|  |
| --- |
| ***The Data Avatar***  I&#39;m Waving at Fat&quot; | Partners in Crime (HD) | Doctor Who - YouTube  When I am first born, I need someone to look after me. I am new and do not really know anything about myself or know what to do. There will be lots of people wanting to look at me and cuddle me and try and teach me things – but it is you that I trust.  I will grow and get bigger both mentally and physically, so I hope you will guide me in what is good and what is bad. If you feed garbage, then I will produce garbage. I will be needy and expensive to keep, but over time I will also bring great value to you.  Remember if you treat me well I will be faithful and trustworthy. You treat me kindly I will do whatever I can to make your life easy, simple, and fun. reat me unkind and I will make your life frustrating.  I will make friends, and I am hoping that your guidance and direction will keep me safe, but sometimes I will meet others who will be different, some good some bad, and I can only hope that the time that you cared for me will make me stay true to my purpose in this world.  Thankyou |

# Data quality considerations

Data is all around us, data is growing exponentially due to technology and the changing expectations of generational population. Despite this, many organisations and government are struggling with how to effectively manage data and quality of the data.

Much has been said about the collection of data, in fact many people refer to the collection of all data. But this is the fundamental problem that has emerged over the limited knowledge of being digital.

Technology solutions have for a long time been seen as the solution to managing data ingestion, storage, use and sharing, but this approach has driven an unprecedented rise in the cost of organisations in managing its data programs. This module will open up the need to pause and re-evaluate this strategy.

It is expected that organisations will collect more data to manage citizens expectations. It is understood that there is a need to do this to simplify interactions and to manage otherwise simple interactions. With this in mind one also needs to consider what is the cost associated with this collection to manage citizen needs.

As a general rule, collecting data for an intended purpose is critical in managing increasing organisational costs.

Managing the collection of right data for a purpose should:

* decrease in cost of architecture
* decrease in cost related to process
* decrease human costs uplifting data maybe for use

Key data terms

When we think of data we need to consider the terms **data, information and knowledge**. When you get a group of people who all have a niche focus on these terms there will be a debate.

To make this simple we will follow some simple definitions:

* **Data:** reinterpretable representation for information in a formalised manner suitable for communication, interpretation, or processing.
* **Information**: is a term used that suggests a richness of meaning and typically takes an end user view of the value of data to an organisation.
* **Knowledge**: suggests an understanding acquired through experience or education that enhances the information.

Another complication in use is structured data and unstructured data. These have been handy tools for marketers promoting software but both forms hide the reality that no data set in digital form is either structured or unstructured.

* **Structured data** contains explicit, discrete elements in tables, columns, keys within a relationship database or tags within a format like XLM.

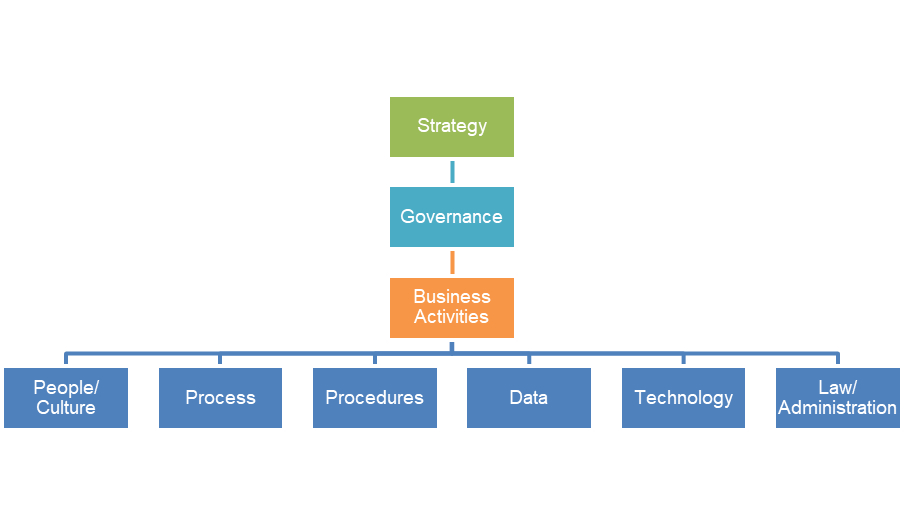
Data in business

* **Unstructured data** are fundamentally text or images which requires human expertise or technology solutions to process the meaning often into a structured format.

Data has fuelled Government decision making, and has help many agencies in understanding natural disasters and in particular recent times with the pandemic. Decision making in these times have been able to assist COVID case locations, testing rates, and which businesses are registering for assistance. This has been a catalyst to generating valuable and insightful discussions and insights by stakeholders to make improvements to citizens lives

Any use of data should support the strategy of an organisation. There should be governance in place to ensure that senior management control and monitor the activity.

A simple organisational structure should flow as follows:



It is the bottom layer of the organisational structure that we will explore in detail to link the process of data quality. Understanding deeply these terms will assist in understanding the maturity level of digitalisation of which data is a subset and as such data quality is a deeper term to learn.

The top level questions of importance are:

|  |  |
| --- | --- |
| People/Culture | * Does my organisation have a strategy that makes reference to digitalisation and the use of data to drive insight to customers? * How purposeful are these statements to what is said and what is reality? * Do senior executives talk digital with meaning, beyond that of what “news snaps” say? * Is there a person responsible in the organisation to showcase this role? * Who are the people accountable to support the person responsible for this function? |
| Process | * Who are the dedicated enabling areas in an organisation designed to bring this to life? |
| Procedures | * Does procedures and practices support the optimised approach to digital and the use of data in the organisation * Is there a central location that guides principles for the organisation |
| Data | * Is data considered an asset, and an understanding that extends to its treatment as an asset |
| Technology | * Is the technology architecture modern and smart to pivot to support the organisations approach to data |
| Law/Administration | * Do laws and administrative practices exist that encourage the collection and use of data to assist citizens to meet obligations |

# What is data quality management?

ISO 8000-2 defines data quality management as:

*Coordinated activities to direct and control an organisation with regard to data quality.*

Data quality management is greater than managing just the quality, it involves the consideration of why data quality is poor and identifying root cause issues. Data quality management is trying to achieve the state of optimised data rather than an ideal state of data perfection.

The Data Management Knowledge of Management (DAMA) practice provides a guide of collective data management techniques to support specific knowledge areas.

These knowledge areas include:

Data architecture;

* Data governance
* Data modelling and design
* Data security
* Document and content management
* Metadata
* **Data quality**
* Data warehousing and business intelligence
* Reference master data;
* Data integrity and interoperability
* Data storage and operations

Data quality management is a critical function of an organisations Data Management Framework. It provides a process for gaining insights into the quality of data in order to make better decisions about improvements to data quality.

Unfortunately the ISO 8000-2requires more effort to understand these definitions. In simple terms the overall approach consisting of different activities to monitor, manage and control data quality with suitable oversight and controls of the activities.

Data quality is also more than just managing data quality. It involves also identifying the considerations of why data is incorrect in the first place.

In general it could be said:

* Data is a key element at the enterprise level;
* By treating data as an enterprise asset, the enterprise can then focus on the value;
* Importantly data quality is the focus on the conformance to a requirement or use rather than an abstract definition of perfection
* Managing data at an enterprise level is critical rather than having a distributed approach where everyone does what they want and how they want.

The virtual cycle of data quality is simple:

Data quality feeds information quality which leads to good decision making and therefore quality outcomes.

There are generally five key principles when consider your approach to data quality management:

1. **Process Centric Approach** – If you are going to address data quality you need to have an enterprise approach to plan to measure, assess, and improve data;
2. **Aim for a marathon to data quality maturity** – while it would be prefect to design a plan, check, and act cycle it really is predicated on the maturity of the organisation. Maturity as a concept is inclusive of everything from technology, data, people/culture, process, and law.
3. **Data specifications** – as best as one can do having standards set in specifications is the ideal state;
4. **Roles and responsibilities** – many people have a role to play in managing data quality, and it is not a single approach reserved just for technology, but is inclusive of ownership and other aspects;
5. **Data quality as a lead** – Data quality is not best led by IT. Ideally an organisation would have a Chief of Data who identifies the co-ordinating role across the data life cycle.

Government policy making is often done with limited understanding of the data that supports policy. Encouraging a better relationship with policy development to be aligned to the data that supports the decision will increase the effectiveness of the measure. Often policy decisions are made with limited understanding of what data will be needed to measure the effectiveness of the measure and subsequently often leads to increased compliance based on poor quality data.

## Data ownership

Organisations often worry about data owners, and assigning data ownership to particular data assets. Organisations ingest, process, use, create and uplift data. It causes owners of data to struggle to retain influence over the activities that change the data including assigning appropriate data quality levels.

An approach to empower owners to maintain explicit specifications for data being created by process, organisations can establish a solid foundation for control of data quality management.

Effective data owners are those that determine the appropriateness of their data for the purpose.

## Data quality management features and benefits

|  |  |
| --- | --- |
| Features | Benefits |
| The organization is aware of the importance of data quality and it is built into all business processes, planning documents and resource decisions. | Quality management decisions are visible and transparent. |
| Data quality policy and supporting processes and procedures are implemented, communicated and followed by all staff. | Consistent data quality management with evidenced based and defendable outcomes. |
| Data quality is a core component of the data lifecycle and is measured and monitored for all key data assets. | Data used is always fit for purpose. |
| Metadata includes data quality which provides transparency and end-user access to data quality metrics. | The quality of data is transparent and metrics reusable. |
| Data quality monitoring and error correction is automated whenever possible. | Data quality management is natural, quick, easy and cheap. |
| Data quality policy is enforced and tested to ensure requirements are met. | Confidence and trust are established when data quality is assured. |

Data quality in the data life cycle

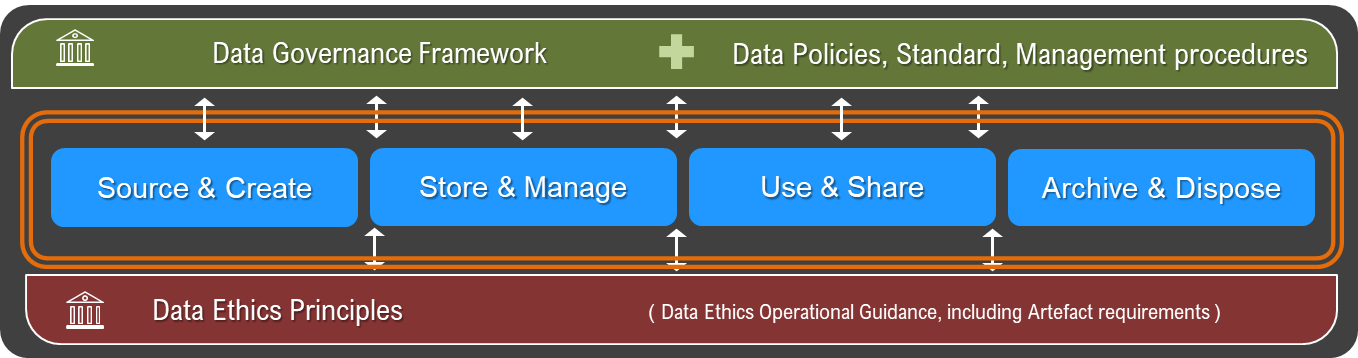
Data like most things has a life cycle. It is the end to end lifecycle of data through 4 stages - source & ingest, manage & integrate, use & share and archive & dispose.

Some modelling looks to utilise an eight-step model expanding out each of the elements above. I would recommend dealing with each of the eight steps as follows:

1. Source;
2. Ingest
3. Manage
4. Integrate
5. Use
6. Share
7. Archive
8. Destroy

Experience illustrates that over time and as your organisation matures you can uplift to the four-step process, but it is important to understand the underlying features of all four steps

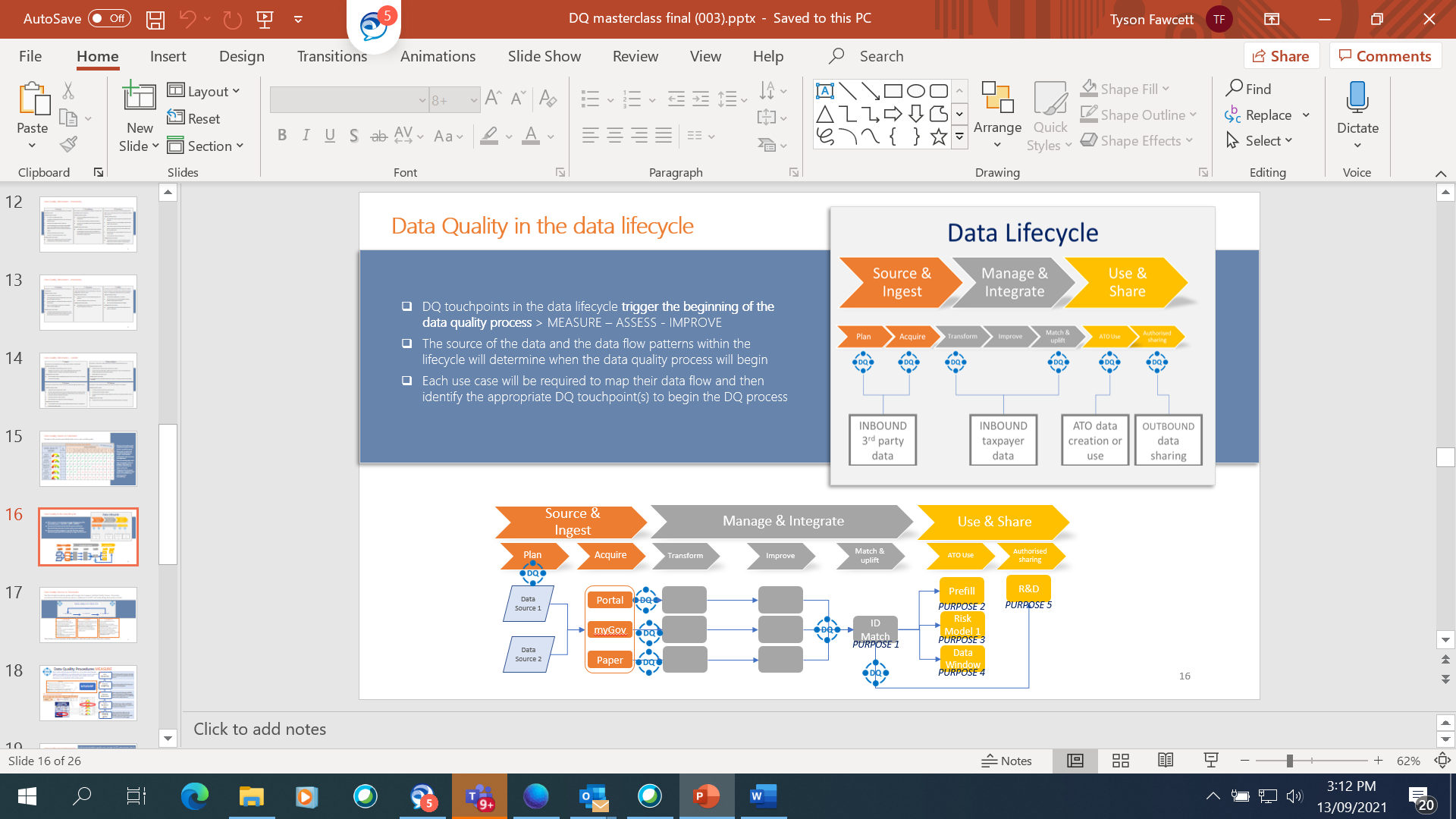
The Australian Taxation Model includes:



A data quality lifecycle covers **the various stages involved when dealing with data**. Data quality management provides a context-specific process for improving the fitness of data that's used for analysis and decision making. The benefits of good data quality management will be delivered in distinct phases throughout the data lifecycle.

It is important to identify **when** the data quality process begins in the data lifecycle. Data may have unique characteristics and/or end to end data flows that vary therefore data quality within the data lifecycle may also vary. Ensure business needs are met when selecting when to begin the data quality process.

In practice:



## Data architecture

Data architecture describes the structure of an organization's logical and physical data assets and data management resources. Data architecture is an offshoot of enterprise architecture.

The concepts of data and technology and their proceeding intersect is a challenge. This is largely because that most business problems have been addressed by way of a technology solution rather than a data solution. When data started to drive accelerated insights in the last decade, it has fundamentally challenged the direction of technology architecture and surfaced a series of technology related improvement opportunities including:

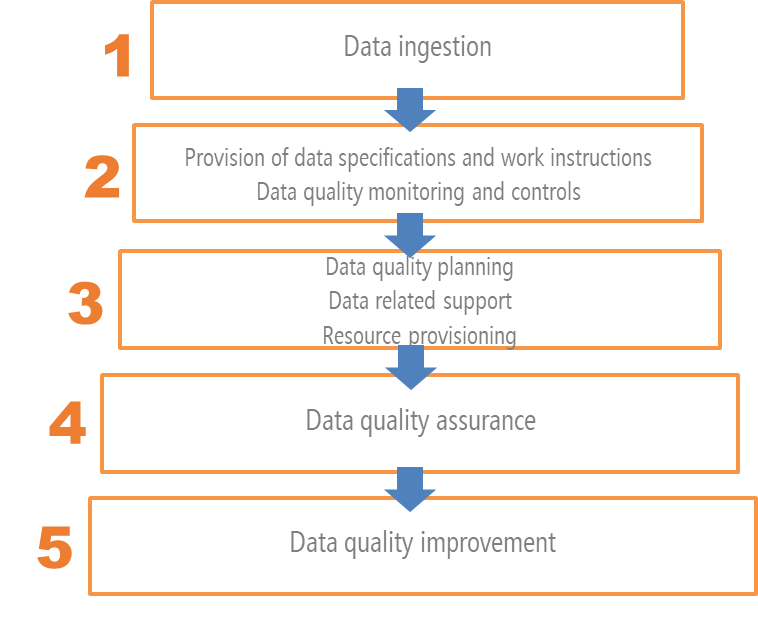
* **Multiple systems** – with more than one entry point for data, and with no effective master data plan, it is hard to determine the correct data value;
* **Complex data architecture** – is where software dictates architecture, which minimises the effectiveness of master data, and also empowers vendor ownership and diminishes organisational control;
* **Data silos** – this is very common where data can not be stored, or the architecture allows for multiple storage locations. The challenge here is managing across data stores particularly the issues of data quality and timeliness and which data element is the golden source.
* **Poor data visualisation** – New software while attractive using poor data will produce poor decision making. Decisions not to manage data quality can be masked in visualisation.

Data quality capability levels

Organisations require a comprehensive data quality capability.

People often refer to a five-level progressive model for data quality management.

It is when all five levels are implemented that one can then say they have a framework that meet data quality management as specified by ISO 8000-61

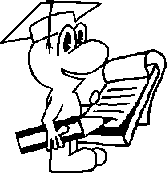


|  |  |
| --- | --- |
| 1 | Ground Zero – this is the level of highly skilled people, who have a deep understanding of the data.  Data quality management doesn’t exist to get gold medals and compliance certificates, but is actually about deriving accurate data fit for purpose, to be delivered to one or more users. |
| 2 | Provisioning – creates additional processes which includes:   * work instructions; * data specifications   The purpose of the provision of data specifications and work instructions is to establish the basis you can compare actual outputs with and the approach to data processing with the intended outputs  A complete scope of this work would include, but not limited to:   * Objectives for performing data processing, thus creating a common standard; * Responsibilities and roles for performing the function * Resources and information needed to perform the function; * Requirements for documenting the production of data, providing an audit trail;   For work instructions and data specifications to be effective one must:   * Involve experts and stakeholders across the organisation; * Ensure that process steps occurred in the correct order * All relevant data attributes are specified; * That data attributes are specified with the correct units of measurement, precision, ranges and allowable values. |
| 3 | Data quality planning;   * Requirements management; * Data quality strategies; * Data quality policy, standards and procedures management; * Data quality implementation planning;   Data related support   * Data architecture management * Data transfer management; * Data operations management; * Data security management   Resourcing   * Data Quality organisational management * Human resource management   With the above ten steps, the purpose of capability 3 is the requirement of management in establishing sufficient understanding of the expectations of stakeholders.  Management will need to cover:   * Creating an overall set of requirements that are appropriate to the approach of data quality management; * Analysis the various inputs from the stakeholders; * Creating a framework to allow to assess individual needs, following an enterprise approach |
| 4 | Data quality assurance   * Review Data Quality Issues; * Provisions of measuring criteria; * Measurement of data quality and process performance; * Evaluation of measurement results   Data quality assurance looks to correspond with the approach *plan, do, check and act.* |
| 5 | Data quality improvements inclusive of:   * Root cause analysis and solution development; * Data cleansing; * Process improvement for data nonconformity prevention |

In summary:

* data quality management addresses the elements of people/culture, process, procedures and practice, data and technology;
* without identifying explicit data requirements no organisation can perform this role;
* the total of the 20 processes of ISO 8000-61 delivers continual, systemic and systemic improvements of data quality

# MODULE 1 ACTIVITY 1-1

1. Explain the importance of data quality.
2. Identify 4 features and benefits of data quality management
3. At what stage in the data life cycle should we consider data quality?

# Conclusion

You should now have a ‘broad brush’ understanding of the main elements of the Data Quality regime. These can be summarised as follows:

* The importance of data quality and its impacts
* Data is key to an enterprise solution
* Treating data as a corporate asset will focus the organisation on delivering value
* Data quality is conformance to a set of enterprise requirements rather than abstract terms
* The process of conformance then allows for solutions directly related to value.

# Module 2: Procedures, process and practices in data quality

## Other resources

You will be referred to the OECD website at <http://www.oecd.org/>, and in particular, the OECD handbook for internationally comparative education statistics.

The following additional resources were also used

* Data Management book of Knowledge (DAMA) - [DMBoK - Data Management Body of Knowledge (dama.org)](https://www.dama.org/cpages/body-of-knowledge)
* Keith Gordon’s book, Principles of Data Management (2013)
* ISO 8000-61 the international standard that specifies a process reference mode for data quality management
* [Elsevier.com](http://elsevier.com/).  <https://www.elsevier.com/books/executing-data-quality-projects/mcgilvray/978-0-12-818015-0>

## Learning outcomes

On completion of this module you should be able to:

* Scope of ISO 8000-61 and its approach to data quality planning data quality control data quality assurance and data quality improvement;
* Understand data quality dimensions and develop indicators, business rules, and apply weightings to measure data quality grades;
* Consider quality standards that dictate permittable uses of data based on quality grades;
* The use of data quality tools to support measurement and when automation of data quality management is beneficial.

## Key concepts

* ISO 8000-61
* Data quality framework
* Data quality dimensions
* Dimension indicators & business rules
* Data quality dimension weighting
* Data quality grades
* Data quality standards
* Data quality tools

# Introduction

This second part of the module expands on data quality and how to develop the practices of an Enterprise Data Quality Framework.

This module includes:

* What is a data quality framework including setting enterprise procedures, processes and policies
* Data quality dimensions, indicators, business rules, and weightings
* Data quality grades and standards
* Data quality tools and when automation of data quality management.

On completion of the module there will be an activity for participants to complete.

This course should be seen as a broad introduction to data quality.

# Data quality framework

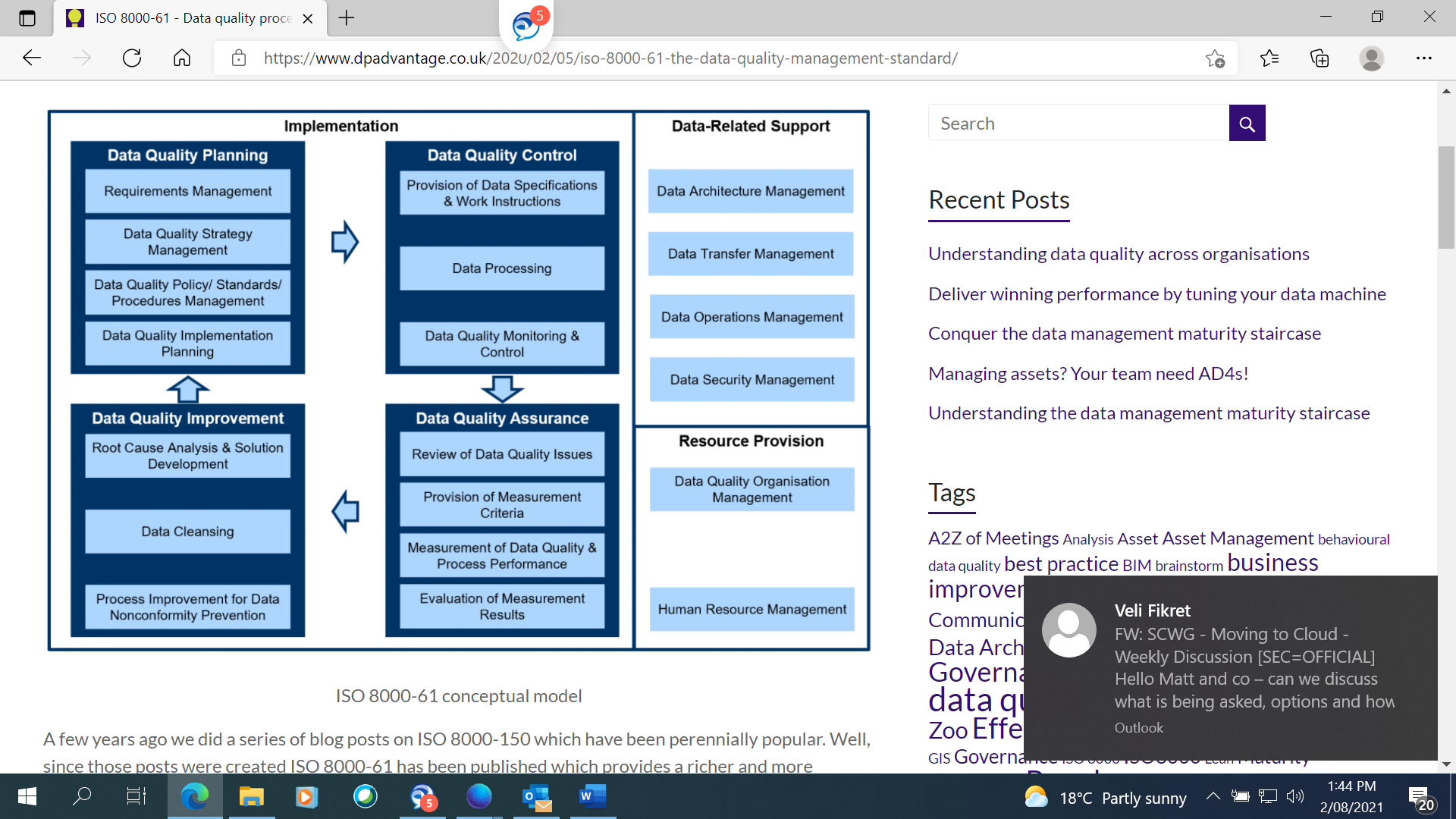
A data quality framework provides a consistent end to end data quality management solution that includes practical components to make it operational and effective.

Benefits of a data quality framework:

* Consistent end to end data quality management supported by data quality specialists
* Ensures data is only used when it is fit for purpose
* Provides foundation to mature the data quality capability and culture
* Roles and responsibilities are clearly defined
* Supports data governance
* Tools and templates are designed to support users through the quality process

#### ISO 8000-61

ISO 8000-61 specifies an overall framework for data quality management. Each of the processes in the ISO can be summarised as:



**Data Quality Control** – Processes to ensure that data arising from activities meets requirements;

* Provision of Data Specifications and Work Instructions – Ensuring data requirements arising from processes are specified, as are the process steps;
* Data Processing – Checking that data arising from processes meets data requirements;
* Data Quality Monitoring and Control – Identify and respond to instances where data processing does not conform to requirements;

**Data Quality Assurance** – This process assesses data quality levels and the performance of processes relating to data quality;

* Review of Data Quality Issues – Assess reported data quality issues to understand their nature and extent;
* Provision of Measurement Criteria – Development of measurement metrics and methods to support data quality measurement;
* Measurement of Data Quality and Process Performance- Engagement of resources to measure data quality levels and assessing the measurement process;
* Evaluation of Measurement Results – Analyse outputs of data quality measurement and assess the impacts of poor data quality and the measurement process;

**Data Quality Improvement** – Deliver sustainable data quality improvements;

* + Root Cause Analysis and Solution Development – Identify root causes of data quality issues and propose solutions to prevent re-occurrence;
  + Data Cleansing – Correction of data quality issues using automated tools and/or human intervention;
  + Process Improvement for Nonconformity Prevention – Implementation of solutions to prevent re-occurrence of data and process non-conformities;

**Data Quality Planning** – Development and agreement of the overall requirements, objectives and plan for delivering the desired maturity of data quality management to the organisation;

* + **Requirements Management** – Identification, definition and prioritisation of the delivery of different data related requirements for the organisation;
  + **Data Quality Strategy Management** – Establish, evaluate and improve the organisations Data Quality Strategy;
  + **Data Quality Policy/ Standards/ Procedures Management** – Development of policies, standards and procedures that support the data quality strategy;
  + **Data Quality Implementation Planning** – Development of a plan that defines roles, responsibilities, sequencing, funding and technology enablers to perform all other data quality management related activities;

The Data-Related Support processes cover:

* **Data Architecture Management** – The overall data architecture of the organisation is understood along with identification of master data sources and approaches to master data management;
* **Data Transfer Management** – Ensure that data transfers meet requirements and that there are suitable audit trails of data transfers;
* **Data Operations Management** – Data environments are effectively managed with effective data processing, capacity planning, suitable backup approaches and management of database technologies;
* **Data Security Management** – Establish and implement data security policies and audit their effectiveness;

# Data quality dimensions

Data quality dimensions are measurable features or characteristics of data used to determine data quality.

|  |  |
| --- | --- |
| **1. Accuracy** | the degree to which the data correctly represents the ‘real-life’ entities they model |
| **2. Completeness** | the degree to which all required data attributes are present and all expected records are present |
| **3. Consistency** | the degree to which data values are consistently represented within a data set, between data sets and consistently associated across data sets. |
| **4. Timeliness** | the degree to which the data is the most up to date version of the information |
| **5. Uniqueness** | no entity or record exists more than once within the dataset, where a key value relates to each unique entity, and only that specific entity, within the dataset |
| **6. Validity** | the degree to which data values are consistent with a defined domain of values e.g. data type, format, precision of expected values, or a specific length of time between data generation |
| **7. Integrity** | the degree to which all references in one table match the values in another table |
| **8. Reasonableness** | the degree to which there is a likely value based on history or other environmental data |
| **9. Currency** | the degree to which the data is current with the real world given likely time related changes |
| **10. Precision** | the degree to which the level of detail in the data element is relevant to the purpose |

# Dimension indicators and business rules

Each data quality dimension requires indicators and business rules to support measuring the quality of data.

**Indicators** describe the relevance of the data quality dimension to the purpose. They define data requirements qualitatively, that is the acceptable levels of data quality required to meet the business needs or purpose of the data.

**Business rules** define the data quality measurement metrics used to support each indicator. They describe the calculations required to determine the dimension score.

Note that there may be more than one indictor or business rule for each dimension in order to calculate data quality score and grade.

#### 1. Accuracy - the degree to which the data correctly represents the ‘real-life’ entities they model

***Determining indicators consider:***

* How well does an attribute reflect reality?
* Do attributes accurately represent the “real world” values they are expected to model?
* Are there incorrect spellings of names or addresses?
* Is there incorrect data that needs to be fixed?

***Developing business rules consider:***

* comparing data to a reliable data source that has been verified as accurate and express as the percentage of accurate records.

#### 2. Completeness - the degree to which all required data attributes are present and all expected records are present

***Determining indicators consider:***

* Does the record fulfill the required expectations of what is comprehensive?
* What is the minimum information essential for your purpose?
* Are there any missing attributes in the record that are critical for your purpose?

***Developing business rules consider:***

* at the attribute, record and/or data set level consider compulsory (possibly legislative) and optional status of data attributes and assign completeness rules at those required levels and express as the percentage of complete records.

#### 3. Consistency - the degree to which data values are consistently represented within a data set, between data sets and consistently associated across data sets.

***Determining indicators consider:***

* Does information stored in one place match relevant data stored elsewhere?
* Are there any instances in which the attributes conflict with itself within or across data assets?
* Data consistency is often associated with data accuracy and may require an assessment of data accuracy of the multiple sources from an additional source.
* Is the data stored across systems, platforms or data assets in sync with each other?
* Are data values the same across the data sets?

***Developing business rules consider:***

* consider format consistency and data standardisation required within and across data sets and express as the percentage of matched values across these various records.

#### 4. Timeliness - the degree to which the data is the most up to date version of the information

***Determining indicators consider:***

* Is the data available when you need it?
* Does the time it takes to obtain data after the transaction or event impact your ability to use the data?
* ‘expected volatility’ i.e. how frequently data is likely to change and for what reasons
* ‘latency’ i.e. the time between when the data was created and when it was made available for use

***Developing business rules consider:***

* consider the maximum time lag between the transaction or event and obtaining the data before it becomes unusable and express as the percentage of records that comply with this time frame

#### 5. Uniqueness - no entity or record exists more than once within the dataset, where a key value relates to each unique entity, and only that specific entity, within the dataset

***Determining indicators consider:***

* Is this the only instance in which this information appears in the data asset?
* Are there any duplicate records in your data asset?

***Developing business rules consider:***

* count duplicate records and express as the percentage of records that are not duplicated in the data asset.

#### 6. Validity - the degree to which data values are consistent with a defined domain of values e.g. data type, format, precision of expected values, or a specific length of time between data generation

***Determining indicators consider:***

* Is it important that attributes follow a set format?
* Does your data have a set of standard definitions or formats required to meet the purpose of the data?
* Is the data in the prescribed format for each attribute?

***Developing business rules consider:***

* count the attributes that meet the required validation rules and express as a percentage.

#### 7. Integrity - the degree to which all references in one table match the values in another table

***Determining indicators consider:***

* Do certain attributes support relationships between data assets?
* If attributes were not present in a record would the relationships between data assets be compromised or unknown? Would this create orphaned or duplicate records?

***Developing business rules consider:***

* Count the attributes required to build relationships between data assets and express as a percentage of records that have integrity.

#### 8. Reasonableness - the degree to which there is a likely value based on history or other environmental data

***Determining indicators consider:***

* Are there maximum or minimum values of data that you can define?
* Are there values that are impossible or unlikely given the data type or source? For example, data that is providing a percentage may not be able to exceed 100.
* Identify outliers and consider accuracy

***Developing business rules consider:***

* count attributes that are within defined value ranges and express as a percentage of records that are deemed reasonable.

#### 9. Currency - the degree to which the data is current with the real world given likely time related changes

***Determining indicators consider:***

* Will the data change over time given its context?
* Is data out of date? Consider attributes where accuracy naturally declines over time for example an attribute that captures ‘age’ will be inaccurate one year after it is captured.
* How often would you expect the data to change over time given its nature? For example, an address within a record that is 20 years old is unlikely to be considered current. A person’s date of birth will not change over time and is therefore always current.
* Is the data up to date? Or how fresh is the data?
* What is the frequency for checking the data value and refreshing it?

***Developing business rules consider:***

* count the attributes that are current as per the defined time frame and express as a percentage of records that are deemed current.

#### 10. Precision - the degree to which the level of detail in the data element is relevant to the purpose

***Determining indicators consider:***

* How precise does the data need to be to meet its purpose?
* Can precision been lost through calculations?
* What level of detail is required to be able to effectively use the data? For example, whole dollars and cents is required.

***Developing business rules consider:***

* Count the records that comply with the level of precision required and express as a percentage.

# Data quality dimension weighting

Data quality dimensions can be weighted in accordance with their relevance to the data purpose using a standard weighting scale. Weighting dimensions allows the data quality score to best reflect the true data quality grade to determine if the data is fit for purpose.

A data quality dimensions weighting scale could be as follows:

|  |  |
| --- | --- |
| **Relevance scale** | **Weighting** |
| **Very high** | **1.00** |
| **High** | **0.80** |
| **Neutral** | **0.60** |
| **Low** | **0.40** |
| **Very Low** | **0.20** |

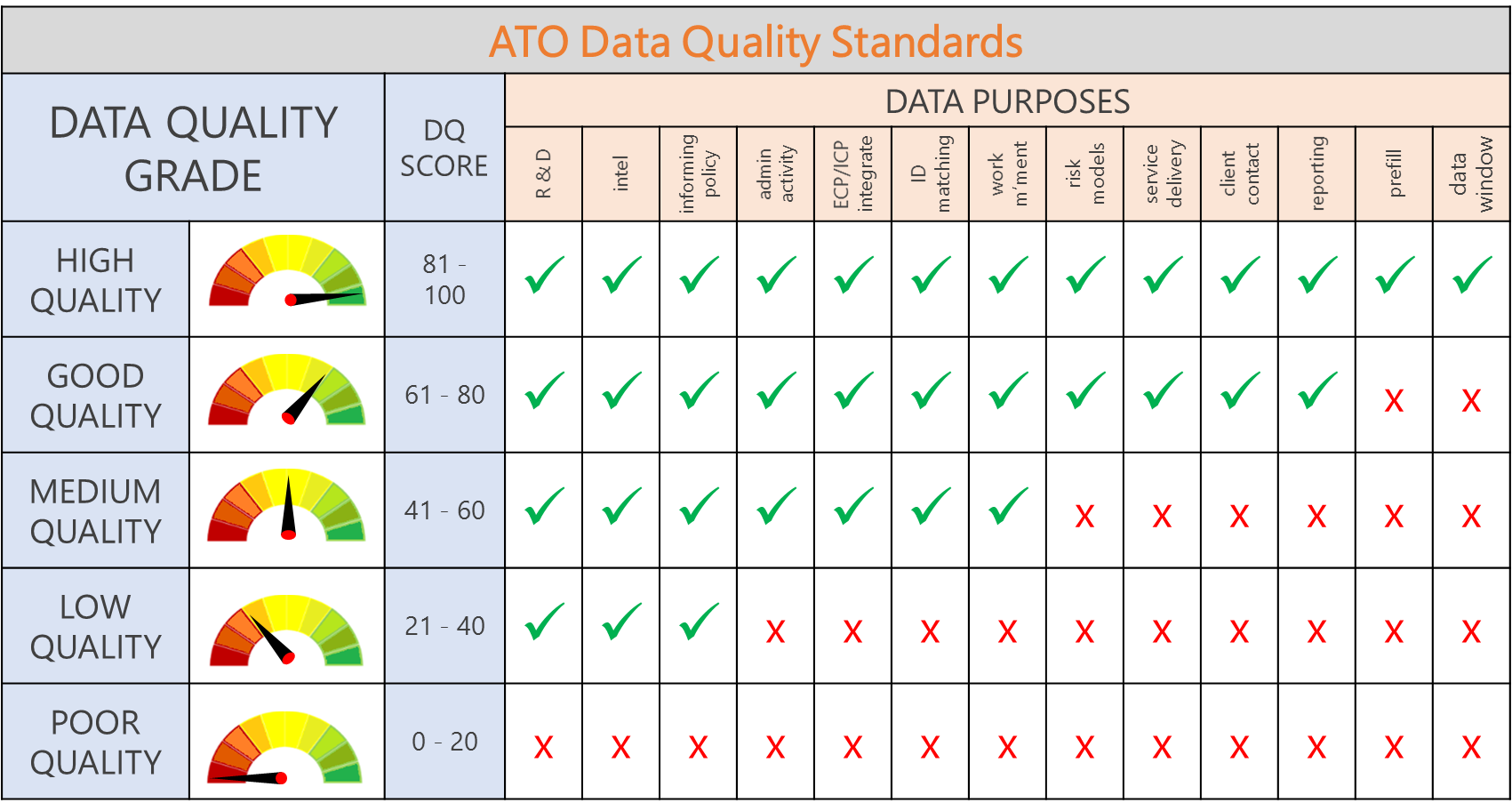
# Data quality grades

The data quality grade is determined using the data quality score calculated at the measure stage in the data quality process.

|  |  |  |
| --- | --- | --- |
| **DATA QUALITY GRADE** | | **DQ SCORE** |
| **HIGH**  **QUALITY** |  | **81 - 100** |
| **GOOD**  **QUALITY** |  | **61 - 80** |
| **MEDIUM**  **QUALITY** |  | **41 - 60** |
| **LOW**  **QUALITY** |  | **21 - 40** |
| **POOR**  **QUALITY** |  | **0 - 20** |

# Data quality standards

Data Quality Standards provide permittable use of data for each data quality grade. Minimum data quality grades are required to be met prior to using the data for specific purposes listed in the standards.



# Data quality matrix

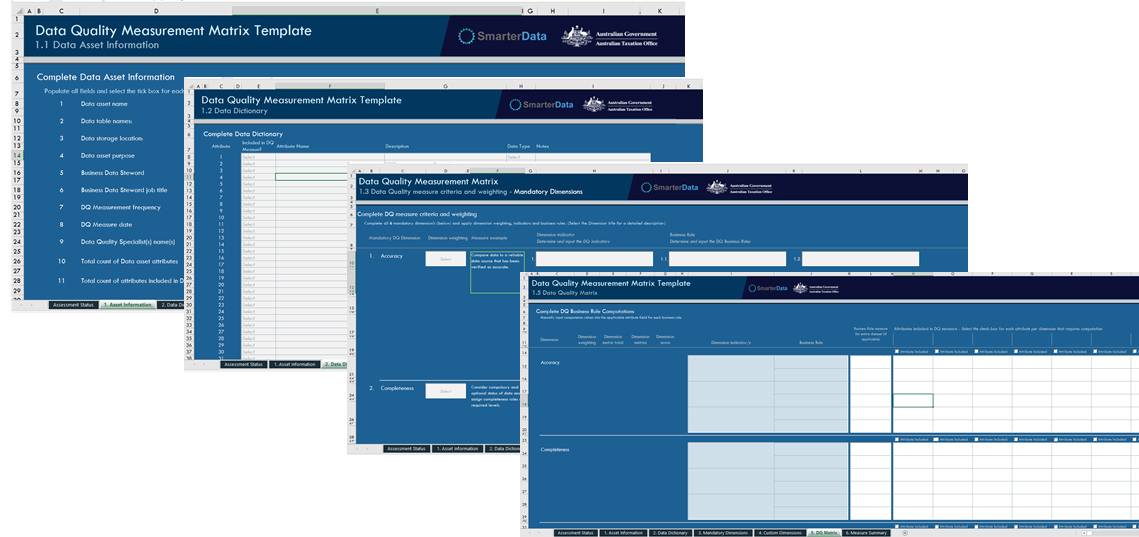
The development of a data quality matrix can help organisations implement consistent data quality management. A self-service **template** that guide users through the required steps to measure the quality of data in order to determine the quality grade of the data asset.

The ATO has developed a matrix template that is a four-step process resulting in a data quality measure report summary that provides an overall data quality grade for the data asset.



The auto features of the template reduce duplication making the data quality measurement process efficient and re-usable over time

Resource intensive at the beginning but efficiencies gained when data quality is monitored and can be re-used by others



Any automation of the DQ measurement process requires the details contained within this template to be determined and recorded, this cannot be avoided. Automation requires human thinking to set the rules which is captured in the matrix.

# Data quality tools

Data quality tools (often software) are often described as the processes and technologies for identifying, understanding and correcting flaws in data. Data quality tools support information governance across operations and decision making.

Data quality software refers to a wide range of tools and services designed to specifically deliver an overall preciseness of data to an organisation. Data quality software solutions offer a broad range of functions and capabilities, which include data cleansing, profiling, parsing, monitoring, and enrichment.

There are several common functionalities in a data quality tool including:

* **Data profiling:** Scanning through data to find patterns, missing values, character sets and other essential characteristics.
* **Data elimination:** The removal of duplicate data and also data that doesn’t meet the desired profile.
* **Data transformation:** For erroneous data that is valuable, it can be transformed into ‘good’ data through correcting typos, standardization, and normalizing numeric values to fall between minimum and maximum values.
* **Data standardization:** Putting data into a common format for easier analysis by standardisation and providing consistency
* **Data harmonization:** Similar to standardization, this practice takes data from a range of sources and puts them into a common format.

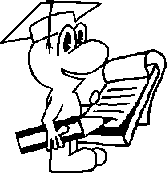
Like the name suggest tools are valuable in moving an otherwise manual process to automation

Gartner provides a Data Quality Solutions Reviews and Ratings to a number of tools for consideration. This can be referenced by [Data Quality Solutions Reviews 2021 | Gartner Peer Insights](https://www.gartner.com/reviews/market/data-quality-solutions)

When selecting a data quality tool one should consider:

* Price – is it one off, enduring, does it include upgrades, and ad on;
* Support – onsite, offsite support, training etc;
* Usability – including IT users, analytics, and business users;
* Scalability – with increasing data ingestion will the tool evolve with that expansion;
* Features – auditing capabilities, integration with existing IT (software and hardware, on previse v’s cloud, meta data support etc

MODULE 2 ACTIVITY 2-1

1. List the data quality dimensions
2. Select two dimensions and write one indicator and one business rule for each.
3. List your organisations data uses and identify the data quality grade that should be required before data is used for that purpose.

# Conclusion

You should now have a ‘broad brush’ understanding of the main elements of the Data Quality regime. These can be summarised as follows:

* The importance of data quality and its impacts
* Data is key to an enterprise solution
* Treating data as a corporate asset will focus the organisation on delivering value
* Data quality is conformance to a set of enterprise requirements rather than abstract terms
* The process of conformance then allows for solutions directly related to value.

# Module 3: Benefits, assessments and challenges

## Other resources

You will be referred to the OECD website at <http://www.oecd.org/>, and in particular, the OECD handbook for internationally comparative education statistics.

The following additional resources were also used

* Data Management book of Knowledge (DAMA) - [DMBoK - Data Management Body of Knowledge (dama.org)](https://www.dama.org/cpages/body-of-knowledge)
* Keith Gordon’s book, Principles of Data Management (2013)
* ISO 8000-61 the international standard that specifies a process reference mode for data quality management

## Learning outcomes

On completion of this module you should be able to:

* Understand benefits of data quality management;
* Assess your data quality management maturity level
* Understand the challenges for data quality management;
* Understand how data quality intersects with data ethics and data value
* Consider data quality issues of the future

## Key concepts

* Digital maturity
* Enterprise data quality management maturity

# Introduction

This final module expands on data quality and its benefits and challenges in introducing it as a core pillar in an organisation.

This is the final part focuses on:

* Benefits of data quality management;
* Assessments of data quality management maturity; and
* Challenges for enterprise data quality management;
* Data quality and ethics and value

On completion of the module there will be an activity for participants to complete.

This course should be seen as a broad introduction to data quality.

# Data quality recap

In the earlier modules you learn that data quality is very important. You will have read that it is a complex initiative, that there is significant effort initially, but as the organisation matures it will shift efforts from manually measuring, assessing and improving to automating aspects of the flow. It will also shift the business model in the organisation to more insights related to data quality across the data management life cycle.

The following information will illustrate the benefits and challenges needed to commence the journey, an enduring and continuous program of data quality.

Finally, I will present a data quality subset of a digital maturity model and suggest some starting points around taking the next steps.

# Benefits of data quality management

In gaining support, particularly from the Executive of an organisation, for an approach to data quality management you are likely to need to provide case studies. One’s that feature a consistent understanding of the type and cost of investment and how that will transform to benefit by data quality management.

This ultimately leads to the question by the Executive:

* What are the benefits of improving data quality management?

In Philip B Crosby’s book Quality is free (1979) is said:

*Data quality is free. It is not a gift… What cost money are the unquality things – all the actions that involve not getting data quality right the first time and all the actions to correct these data quality issues…*

At a very high-level data quality benefits will include:

* Organisational efficiencies and costs associated with this;
* A more robust risk management approach as defined by better quality data;
* The ability to explore better and new opportunities

## Organisational Efficiencies

It has long been a strategy for organisations to collect lots of data and not necessarily the right data. When *big data* become a word of populist approach, immediately all data become important to collect.

This led to many organisations consuming lots of data and subsequently increasing the overall costs associated with the management of data. This approach has matured pivoting from collecting all data to only collecting data that is fit for purpose.

Poor quality data increases costs of:

* storage where data is not used;
* resources to cleanse data before use;
* processing
* governance to manage
* security
* extra effort across storage

Benefits of a data quality framework:

* reduce expensive and duplicate data cleansing/uplift activities by identifying and addressing the root cause of poor-quality data at source where applicable
* provides quality assured data sets for downstream use
* uplifts the compliance risk management framework with quality improved data used to mitigate risk
* reduces internal infrastructure costs by only collecting and improving the right data
* improves decision making re investments to improve data quality
* reduces the cost of managing our data
* provides transparency of DQ measurements that are reusable across the data lifecycle
* increases the value of data
* increases the integrity and trust in data

## Risk management

Data quality management goal is to determine the appropriateness of the data for its purpose.

Risk management relies on data to measure and asses organisational risks including modelling and sharing. By improving the quality of data you should expect that modelling and sharing should become better.

In addition, it should also provide better visibility to performance metrics of the data.

Good quality data should:

* Make for better decisions;
* Better target assurance and compliance approaches;
* Provide an organisation with an accurate understanding of its data holdings;
* Provide users trust in the data being used;

## Explore new opportunities

Having good quality data will allow for deep exploration work to be performed by data scientists. Data scientists uncover the answers to major questions that help organisations make objective decisions.

Data scientists rely on good quality data to solve complex problems through trend and pattern analysis, customer experience engagement, predictive modelling, statistics and analytics.

# Preparing for data quality

Implementing a data quality practice is like any other program of work designed to improve an outcome. One must set a vision, followed by a goal, establish your objectives, plan the execution, execute the plan, and finally evaluate the performance.

Some preparatory activities for an organisation are:

* Agreeing the vision, goals and objectives of implementation
* Stakeholder identification and engagements;
* Executive support;
* Planning;
* Budget and resourcing agreement;
* Implementation approach;
* Aligning to synergies of other projects
* Identifying and reducing risks

Data quality is only effective when coherent with drivers of change in an organisation. Data quality organisational change is often stated in:

* Organisational planning documents;
* Visionary leaders with an understanding of its investment and benefits;
* Skilled capability;
* Key internal and external stakeholder groups that understand the importance of data quality.

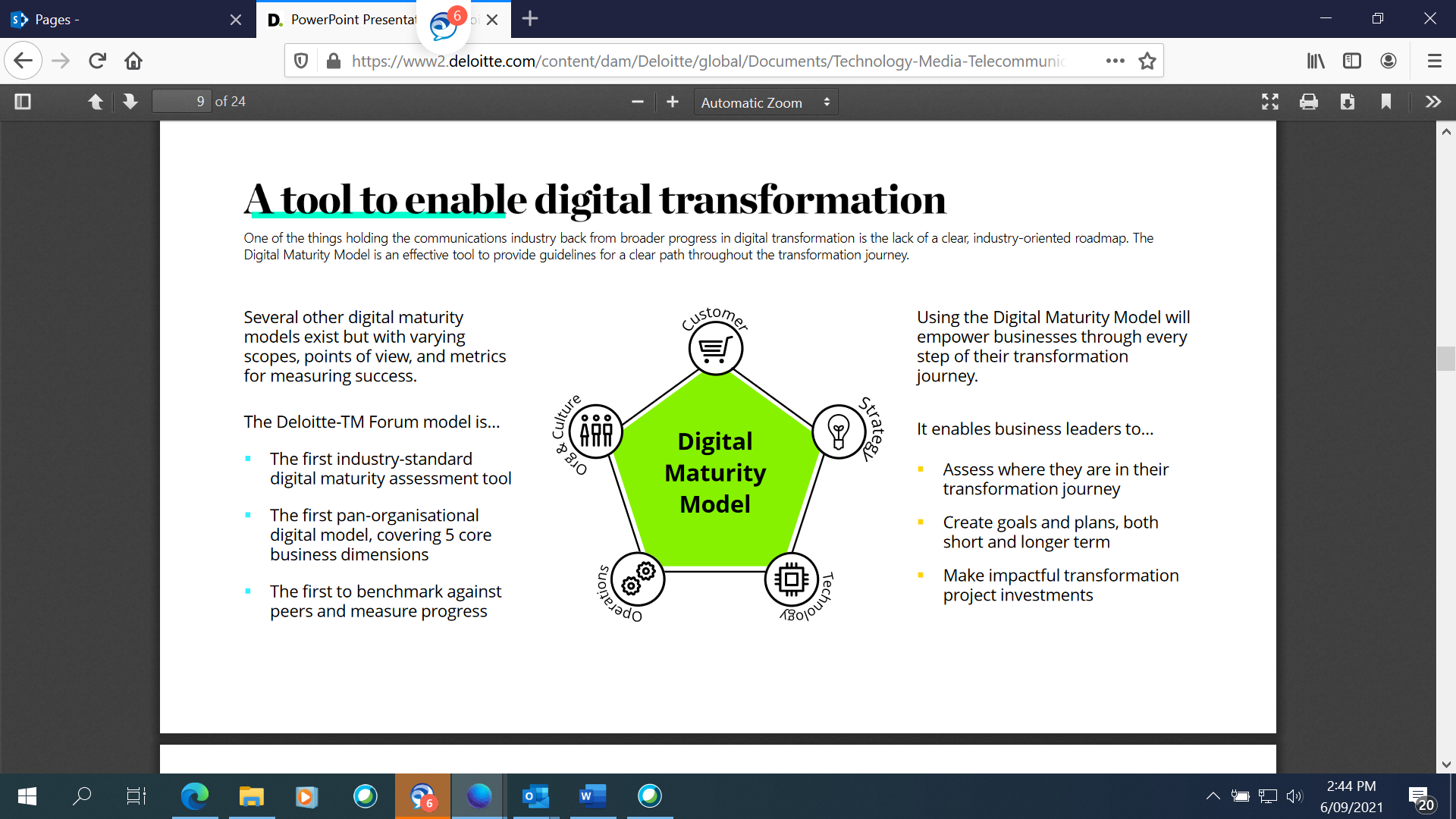
# Enterprise maturity assessments

Many people would be familiar with a digital maturity assessment. They commonly consider all aspects of an organisation using themes related to digital with scales to assess an organisations current state.

There are many digital maturity assessments available for use. It is recommended that you find one that best suits your organisation by using a comparative company.

#### Digital maturity assessments

Some examples and features of digital maturity assessments include:





Essentially the most common themes used in a digital maturity assessment are:

1. Governance and leadership;
2. People and culture;
3. Capacity and capability
4. Innovation
5. Technology

Scales to measure each theme include:

Level 1: Minimal;

Level 2: Informal and reactive;

Level 3: Transitional

Level 4: Customer Driven;

Level 5: Transformative

With each level having a set of descriptions to describe a sense of measurement. It is measured on a basis point of 1 ♦1.5 ♦2 ♦2.5 ♦3 ♦3.5 ♦4 ♦4.5 ♦5 (highest)

The DMBOK outlines nine key practice areas (KPA) in data management. While one would measure all the elements, this is a study guided focused on Data Quality and should be the focus of measurement

* Data Governance
* Data Development
* Database Operations Management
* Data Security Management
* Reference & Master Data Management
* Data Warehousing & Business Intelligence Management
* Document & Content Management
* Metadata Management
* **Data Quality Management**
* Data Architecture Management

The five maturity levels used by Mrs. Reeve (the Carnegie Mellon original CMM names are in parenthesis) are:

|  |  |
| --- | --- |
| 1. Immature   (Initial) | The best practice activities are not performed by the organization.  The best practice tools are not available or not used. |
| 1. Repeatable   (Repeatable) | Some parts of organization are using recommended tools and processes while other parts are not. |
| 1. Managed   (Defined) | The organization has a documented standard for performing the assessed activity or activities consistently and using applicable tools effectively. |
| 1. Monitored   (Managed) | The process in question is established, tracked and monitored. Recommended tools are in place and being used consistently across the organization. |
| 1. Continuous Improvement   (Optimizing) | The activity is continually reassessed, improved upon, tracked and built into process. |

#### The Questionnaire – Activities

The first stage is to identify the data managment areas to be assessed, the stakeholders to interview, and the questions to be asked as part of the interviews/assessment.

The second stage is conducting the assessment interviews with questions. Below are some sample questsions, covering aspects of data qiuality

* What authority/legislation/agreement was the data collected under?
* Which organisation(s) compile the data,
* About whom, or what, was the data collected?
* What key data items are available?
* What was the original purpose for collecting the data?
* What does the data not represent or cover?
* Have standard classifications been used?
* How often is the data collected or expected to be collected?
* When did the data become available?
* What is the reference period for the data?
* How was the data collected?
* Has the data been adjusted in any way? If so, how much was adjusted and on what data items?
* What is the collection size?
* What are the standard errors for the key data items?
* What steps have been taken to minimise processing errors?
* How consistent is the data over time? If there are differences, what are they and what is their impact?
* Is a time series available for this data?
* Have there been changes to the underlying data collection?
* What other data sources is this data comparable with?
* What other data sources in society report similar information? How do these data
* Is there a particular context that this data needs to be considered within?
* What other information is available to help users better understand this data source?
* Are there any ambiguous or technical terms that may need further explanation?  
  What are the contact details for requesting more information?
* In which formats is the data available for people to use? Where and how do you access them?
* Are there any privacy or confidentiality issues that prevent the data from being released publicly?

For this example, some general operational priniciples to consider are:

* Interviewing a cross section of people will be important; and
* Not all interviewed will answer all questions;

A wide range of stakeholders should be engaged

The most valuable perspectives are those from the business people who rely on the organizational data management capabilities.

#### Data Quality Maturity Range

The Data Quality Maturity Range looks to assess across several layers. One of the better examples I have seen looks to assess components across:

1. **Data Quality Expectations** - defining data quality expectations often is tied to the convergence of understanding between the information specialists and their business clients.
2. **Dimensions of Data Quality** - the ability to predict where data quality becomes critical to achieving business
3. **Policies** - the approach to managing conformance
4. **Procedures** - high-performance organization lies in its well-defined processes and protocols for ensuring data quality
5. **Governance** – is mapped program to the different levels of the maturity model;
6. **Standards** - is a key to coordinated activities, and the maturity of the organization is reflected in the way it defines and implements data standards;
7. **Technology** - the focus on data quality improvement transitions from acquiring tools to assembling a service-oriented approach.

Use descriptions for each maturity level to describe the component state. Once mapped it should follow that it will illustrate where improvements can be made.

For example:

|  |  |
| --- | --- |
| 1. **Data Quality Expectations - component maturity description** | |
| Initial | * Data quality activity is reactive * No capability for identifying data quality expectations * No data quality expectations have been documented |
| Repeatable | * Limited anticipation of certain data issues * Expectations associated with intrinsic dimensions of data quality (see chapter 8) associated with * data values can be articulated * Simple errors are identified and reported |
| Defined | * Dimensions of data quality are identified and documented * Expectations associated with dimensions of data quality associated with data values, formats, * and semantics can be articulated using data quality rules * Capability for validation of data using defined data quality rules * Methods for assessing business impact explored |
| Managed | * Data validity is inspected and monitored in process * Business impact analysis of data flaws is common * Results of impact analysis factored into prioritization of managing expectation conformance * Data quality assessments of data sets performed on cyclic schedule |
| Optimized | * Data quality benchmarks defined * Observance of data quality expectations tied to individual performance targets * Industry proficiency levels are used for anticipating and setting improvement goals * Controls for data validation integrated into business processes |

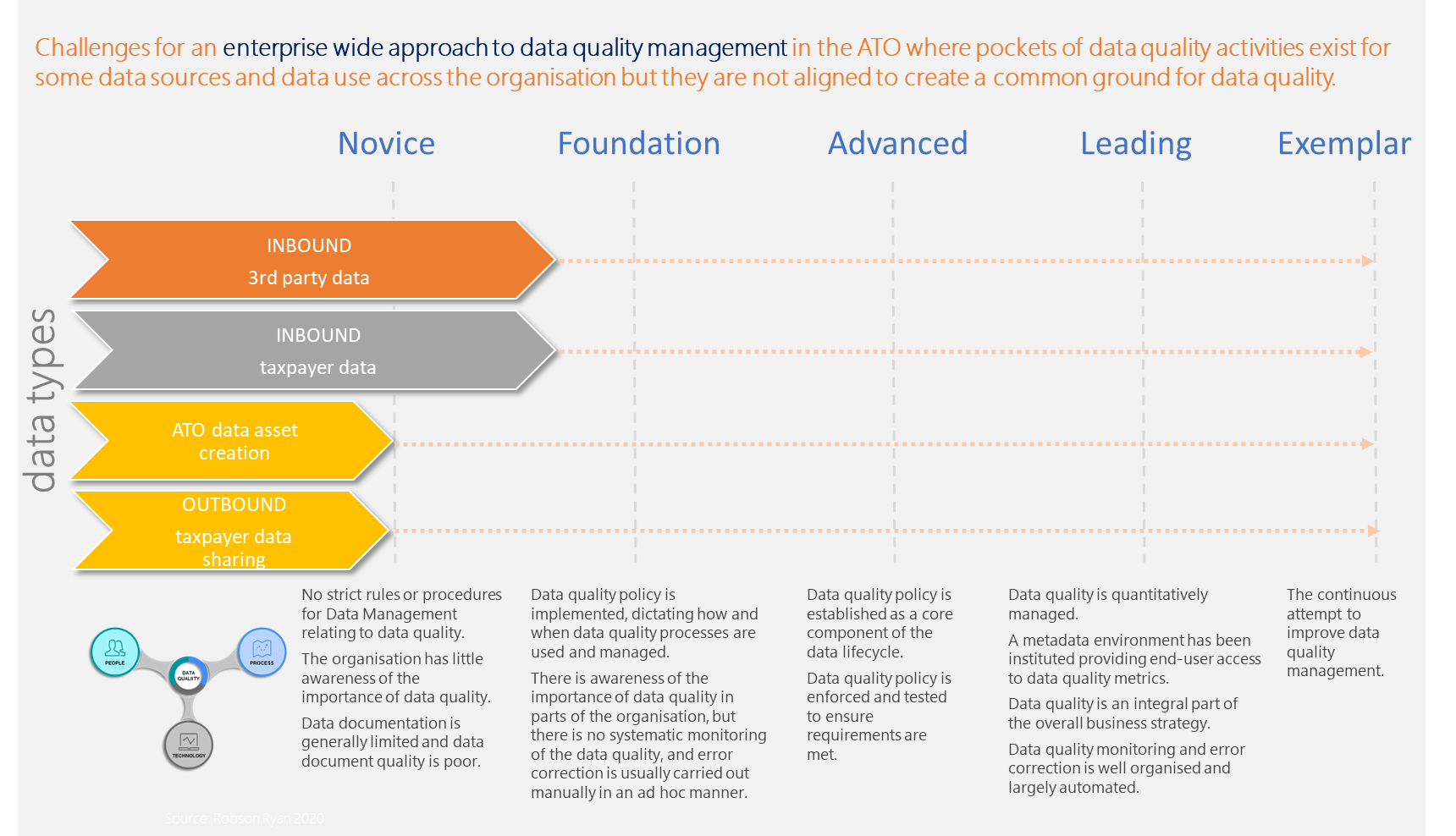
#### Data quality management maturity assessment

* Building a DQ Management Maturity Assessments: to understand your level of maturity, where to prioritise resources to uplift in people, process, procedures, law and technology

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Questions to consider** | **Novice** | **Foundation** | **Advanced** | **Leading** | **Exemplar** |
| People/ Culture | What is the level of awareness of DQ in your organization?  How important is DQ to your organisations?  Does your organization have roles dedicated to DQ management? | The organisation has little awareness of the importance of data quality.  Data leaders cannot identify the overall quality of ATO data at any given point in time. | There is awareness of the importance of data quality in parts of the ATO  Dedicated DQ operational roles are established for ongoing BAU DQ management. | DQ is a core KPI - reported regularly and consistently.  Leaders can identify the quality of data at any time. | Data quality is an integral part of the overall business strategy and planning processes. | The continuous attempt to improve the data quality capability. |
| Process | Are DQ issues addressed downstream by multiple data users?  Is DQ built into business processes as BAU?  Do you consider the quality of data before using it in your organisation?  Does your organisation have KPI’s linked to the improvement of DQ?  Are any DQ measurements/assessments checked for accuracy?  Does your data governance function include DQ?  Does your enterprise risk management include DQ risks?  Is your DQ process automated where applicable? | DQ issues are addressed downstream multiple times by multiple users – duplication of effort  Data that is of poor quality is not prohibited from use  Data quality risks are not identified measured or monitored consistently across the ATO. | Data quality policy is implemented, dictating how and when data quality processes are used and managed.  The DQ process is manual at first with automation developing for DQ measurement, DQ reporting and DQ uplifting  Data improvements will be funded to address root cause only where the long-term benefit outweighs the cost. | Data quality operations is BAU  Data quality policy is established as a core component of the data lifecycle.  Data quality policy is enforced and tested to ensure requirements are met.  DQ risks are identified, measured, managed and monitored as regular part of business management. | Data quality monitoring and error correction is well organised and largely automated.  DQ metrics are captured in metadata in an enterprise data catalogue  Data quality policy is enforced and tested to ensure requirements are met - governance. | The continuous attempt to improve data quality management. |
| Procedures | Does your organisation only consider DQ when issues arise?  Does your organisation have DQ policy, process and procedures?  Is DQ measured or monitored regularly?  Is DQ measured consistently across your organisation?  Is DQ quantitatively managed?  Are DQ procedures easy for staff to follow?  Are DQ results and decisions recorded and made transparent to data users?  Does your enterprise metadata include DQ? | No strict rules or procedures for Data Management relating to data quality.  Often data quality is reactive - only considered when data issues arise.  DQ assessments are subjective, challengeable and ineffective for end to end lifecycle data management.  Data documentation is generally limited, and quality is poor. | There is no systematic monitoring of the data quality, and error correction is usually carried out manually in an ad hoc manner. | Procedures are easy, natural | Data quality is quantitatively managed.  A metadata environment has been instituted providing end-user access to data quality metrics. | The continuous attempt to improve data quality procedures. |
| Law/ Administration | Does your organisational laws and administration allow data to be corrected for a variety of data types | No changes can be made | No changes can be made and data can only be changed by the source | Partial changes can be made by way of verified data sources | Data can be corrected and changed and notifies the source of the changes as permitted by law | Data correct and shared with permission by the source to other organisations |
| Technology | Is your DQ process managed manually?  Are any parts of the DQ process automated? For example, measure or improve.  Does your organisation have technology that is dedicated to managing data quality? | No technology support DQ management | DQ technology is emerging but often not used consistently or effectively across the organisation to manage DQ | Data quality monitoring and error correction is automated whenever possible.  DQ technology develops to automate DQ where applicable in the process. | DQ tools provide fully automated solutions in the DQ process | The continuous attempt to improve data quality tools and supporting technology |

#### Enterprise data quality management maturity

Data quality management maturity may differ across the organisation based on the source and/or characteristics of the data used in your organisation.



# 

# Challenges for data quality management

The minute that data is ingested by an organisation, the quality of that data is at risk. As data flows through systems, processes and various environments, its integrity is consistently threatened. This leads to operational risks.

An operational risk largely aligns to data use, whether that be inaccurate and misinformed decisions, sending out incorrect communications to customers or even regulatory noncompliance if data is not fit for purpose.

The challenges of managing enterprise data quality can be summarised as:

* With many data stores across various domains it can be difficult to align to a common standard;
* As data stores expand the process of standardisation becomes more difficult to align;
* Ownership and stewardship can be weak because of the unknown;
* There can be many different software solutions that include their own data stores and/or undertake uplifting data;
* Business processes are updated by different parts of the organisation;
* There can be many users of the data, with various understandings of the purpose of the data;
* Processes are not centralised to the data store or the source of truth;
* The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration;
* Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.
* Data change very fast and the “timeliness” of data is very short, which necessitates higher requirements for processing technology.
* No unified and approved data quality standards;

# Data quality and data ethics

Data ethics is a fast maturing discipline that aims to address the pace of technology changes and processes which ultimately are not covered by slow moving legislation.

Just because you can do something with data does not mean you should do it

Many organisations these days are developing:

* Data ethic frameworks;
* Data ethic principles
* Data ethic policies; and
* Appointing data ethical officers

The pillars of modern tax administration system are:

*Stability – The central purpose of taxation is to fund Government expenditure on public services. In order to fulfil this purpose a tax must be sustainable*

*Efficiency – An efficient taxation system minimises the distortionary effects and unnecessary influences of taxes on the behaviour of consumers and producers*

*Equity – In-principle, taxes should be both horizontally and vertically equitable. Horizontally equitable taxes tax people in similar financial circumstances in the same way. Vertically equitable taxes are progressive, imposing higher taxes on individuals with greater capacity to pay*

*Simplicity – Taxes should be simple, transparent, practical and enforceable, with minimal administration and compliance costs*

*Competitiveness – Taxes should be aimed at improving competitiveness of Australian businesses both domestically and internationally*

*Revenue adequacy – Tax reform measures should aim to minimise significant impacts to the economy by avoiding sudden large-scale expenditure cuts.*

Using poor quality data and making decision impacts all the above features. The most significant impact is equity when you consider that some taxpayers could be selected for review by an administration in the absence of having quality data that considers the entire population.

*In simple terms poor quality data makes for poor quality decisions*

While it is an important to commence some understanding of ethics the study of what will work in an organisation needs to be considered in the context of the maturity of your business.

# Data quality and data value

Key to data quality is the direct data value that it generates. Previously we have identified a number of benefits established by data quality.

Data quality and its value:

* promotes informed decision-making
* business decision, at the individual level or corporate level, has an indirect or direct impact on the bottom-line.
* decisions driven by poor data practices, causes more harm than good.

*84% of CEOs have*[*conveyed concern*](https://www.forbes.com/sites/forbesinsights/2017/06/05/the-importance-of-data-quality-good-bad-or-ugly/#5ea2ae5910c4)*over the lack of trust in the quality of the data while making decisions.*

* quality standards of good data, organizations can develop better control over the outcomes of these decisions. This improves confidence in data, efficiency, eradicates errors and lowers risk.
* improves customer experience

Ensuring high data quality leads to better customer engagement. In their tech blog, Netflix wrote at length about their data strategy and how it has helped them to improve the personalization and recommendation engine of their business.

*“One thing we have found at Netflix is that with the great availability of data, both in quantity and types, a thoughtful approach is required to model selection, training, and testing. We use all sorts of machine learning approaches: from unsupervised methods such as clustering algorithms to a number of supervised classifiers that have shown optimal results in various contexts”*

Netflix’s estimates the combined effect of personalization and recommendations save them more than $1B per year. The high quality data allows Netflix to help the company improve their:

* + take-up rate;
  + overall engagement; and
  + reduction in subscription cancellation rates.

# Data quality issues of the future

Organisations are dynamic and exist in dynamic environments. If the data environment was a steady state, then there would be no need to evolve thinking on aspects of data quality.

However, the above is not reality and as such sources of data change, technology has accelerated, and customers are international. In addressing these changes, the challenges of data quality will be influenced by:

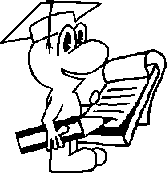
* changing organisation structures;
* changing markets and supplier relationships;
* new legislation;
* changes to external data feeds or requirements; and
* internal software changes.

Data quality in the future will be as sound as quality management was in the 70s. The concept of data quality will modernise with the inclusion of technology as things progress. The single biggest risk in the future to data quality is the dilution of human knowledge in the process, procedure and practices of data quality with the illation of technology being the fix.

The critical factor for data quality in the future is to imbed specialisation in data quality within organisation as a core to future data changes. This specialisation team can be used to:

* establish the data quality framework for the organisation;
* have skills to support data quality methodologies;
* act as specialised advisory team across the data life cycle on data quality; and
* assist in automating aspects of the data quality within the data quality life cycle

# MODULE 3 ACTIVITY 3–1

1. How would you rate your own organisations data quality knowledge?
2. What would be your assessment of your organisations data quality management maturity level?
3. What do you think will be some of the challenges in implementing a data quality framework in your organisation?

# Conclusion

You should have some ideas on the benefits, challenges and the starting point to do a data quality maturity assessment to identify your strengths and prioritisation gaps in your organisations.

Good luck

# GLOSSARY

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Accuracy | A data quality dimension. The degree to which the data correctly represents the ‘real-life’ entities they model. |
| ATO data quality capability | All staff who manage data quality. |
| ATO Data quality framework | Components that together provide the end to end data quality management solution. |
| Data quality strategy | An organisations plan and commitment to maturity the data quality capability. |
| Completeness | A data quality dimension. The degree to which all required data attributes are present and all expected records are present |
| Consistency | A data quality dimension. The degree to which data values are consistently represented within a data set, between data sets and consistently associated across data sets. |
| Currency | A data quality dimension. The degree to which the data is current with the real world given likely time related changes. |
| Data governance activities | Activities conducted to independently check that data management policies are being adhered to. |
| Data life cycle | The end to end lifecycle of data through 4 stages: source & ingest, manage & integrate, use & share and archive & dispose. |
| Data quality | The extent to which data is fit for purpose. |
| Data quality activities | Activities that measure, analyse and improve data quality. |
| Data Quality Assurance | This process assesses data quality levels and the performance of processes relating to data quality |
| Data Quality Control | Processes to ensure that data arising from activities meets requirements |
| Data quality dimension | A characteristic of data that can be measured. Standard DQ dimensions apply to all data. Custom DQ dimensions are selected If they are relevant to the data and its purpose/use. |
| Data quality dimension business rule | The business rule describes how each data quality dimension indicator is measured. |
| Data quality dimension indicator | Indicators describe the relevance of the data quality dimension to the purpose. |
| Data quality dimension weighting | The weighting given to each data quality dimension based on its relevance to the purpose/use of the data. |
| Data quality dimension weighting scale | The scale used to apply a data quality dimension weighting. |
| Data quality grade | The grade applicable to the data quality score. It can be categorised as high, good, acceptable, low or poor. |
| Data quality improvement | A set of activities to improve the quality of data. |
| Data quality management | The end to end management of the data quality framework. |
| Data quality measurement | The first step in the data quality process is data quality measurement. |
| Data quality measurement criteria | Established for each data quality dimension providing the details of how the dimension is measured consisting of dimension indicators and indicator business rules. |
| Data quality metrics | The measurements by which you asses the quality of data. |
| Data quality monitoring | Is the continual assessment of data quality dimensions to determine if previously identified data issues have been addressed. |
| Data Quality Planning | Development and agreement of the overall requirements, objectives and plan for delivering the desired maturity of data quality management to the organisation |
| Data quality procedures | Detailed procedures that guide users through the three stages of the data quality process – measure, assess and improve. |
| Data quality process | Ongoing process consisting of three stages – measure, assess and improve. |
| Data quality score | The cumulative score of the measurement of each data quality dimension. |
| Data quality standards | Dictate permittable data uses for each data quality grade. |
| Fit for purpose | Data that is of sufficient quality to be used for its intended purpose. |
| Good quality | A data quality grade used to describe a data asset with a data quality score between 61 and 80. |
| High quality | A data quality grade used to describe a data asset with a data quality score between 81 and 100. |
| Integrity | A data quality dimension. The degree to which all references in one table match the values in another table. |
| Low quality | A data quality grade used to describe a data asset with a data quality score between 21 and 40. |
| Medium quality | A data quality grade used to describe a data asset with a data quality score between 41 and 60. |
| Metadata | Information about the data quality measurement of data asset. |
| Poor quality | A data quality grade used to describe a data asset with a data quality score between 0 and 20. |
| Precision | A data quality dimension. The degree to which the level of detail in the data element is relevant to the purpose |
| Reasonableness | A data quality dimension. The degree to which there is a likely value based on history or other environmental data. |
| Relevancy | A data quality dimension. The degree to which the data is relevant to business operation or performance |
| Root cause | The fundamental reason for data quality issues. |
| Timeliness | A data quality dimension. The degree to which the data is the most up to date version of the information |
| Uniqueness | A data quality dimension. No entity exists more than once within the dataset, where a key value relates to each unique entity, and only that specific entity, within the dataset |
| Validity | A data quality dimension. The degree to which data values are consistent with a defined domain of values e.g. data type, format, precision of expected values, or a specific length of time between data generation |