Data Management – Guide

(Source: free.itmatters.com.au)
Version 1.1 – Release date:09 February 2021

Our focus is to help you improve the management of your data to benefit your business, community group, and/or personal well-being.

Terms of Use
The intent of each guide is to give you a starting point in the activity addressed by the guide, whether to address the data management needs of yourself as an individual, or the more complex needs of your business, or your community group.
Each guide intends to give you guidance for your data management project so you understand what needs to be addressed at a high level and which skills are required to conduct your data management project so you can either acquire such skills or enter in an agreement with a data management professional.

The examples are provided with rationale. These rationale may not be relevant for your needs, and are not meant to be either right or wrong as all projects are different; they are just examples of what could be and documentation of why such decisions are made. It is important to **develop the habit of recording rationale** as these can be reviewed over time and allow old decisions to be re-assessed and new decisions to be made. **We strongly advise to document a rationale alongside each decision.**

Assumptions
It is assumed that you have already become familiar with some common data issues, that you have realised how they affect your life – whether at a personal or business level, and that you are now seeking to understand what may be involved to address such issues.

Version History

Ref.	Version Date	Version Details
v1.1	09 Feb 2021	Minor changes
v1.0	28 Jan 2021	The initial release provides examples for the needs of the individuals. Examples for Community Groups, and Businesses are still to be fully released.

Purpose

This guide intends to provide a brief explanation on the eleven knowledge areas identified in the **professional framework developed by DAMA** for performing data management.

Such professional framework is **used as a structure to guide the explanations** on the management of data. Such explanations aim at helping you decide on your priorities in terms of addressing your data issues.

Given that this guide does not assume any technical background, the content is addressing personal data issues, as well as community, or business data issues, with a focus on conveying meaning rather than academic exactitude – drawing on examples and accumulated knowledge from over 20 years of consulting in the data management area mainly as a data modeller and educator – with corporations such as Oracle Corporation and IBM, and a wide variety of government agencies – mainly Australian Commonwealth.

An example will be used for each of the targetted audience group (ie individuals, community groups, and businesses) and such example will be explored in the eleven knowledge areas at stake in order to help you develop a sense of the process you will need to go through for your other data needs.

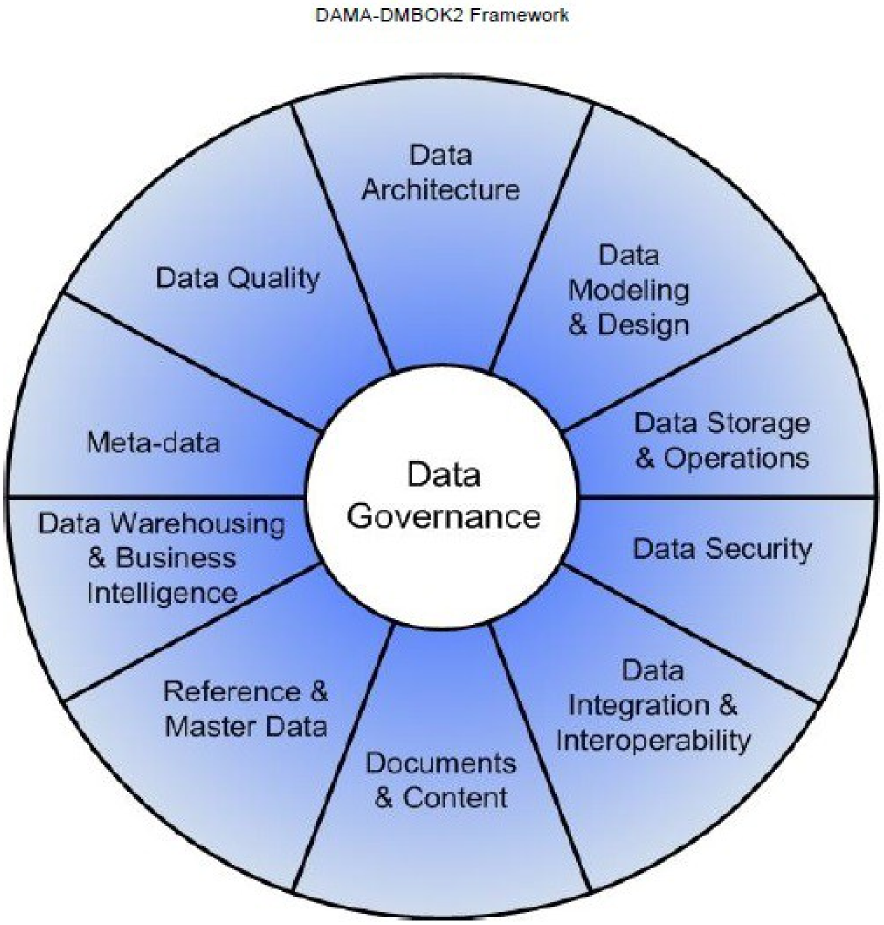

Figure 1. The DAMA-DMBOK2 Guide Knowledge Area Wheel

Table of Contents

Data Governance

Data Governance is the **core activity** whose output includes the **decisions** / policies necessary to manage the data in a way which is likely to better benefit the individual, community group, or business at stake.

The decision making group is also expected to **enable** the management of the data according to the agreed decisions / policies, and **manage** the implementation of the policies – through reporting, controls, and training.

With Data Governance in mind, the commitment is to **manage data as an asset** in the same manner as physical items (cars, parts etc), human resources, and finances are managed and accounted for.

Data Governance provides a central focus for identifying and controlling the collection, storage and retrieval for usage or reporting, update, and disposal of data as an asset.
This data cycle is often known as **CRUDA**: Create-Retrieve-Update-Delete-Archive.

Data Governance for Individuals

<u>Example of decisions / policies</u>:

- My health data is to be managed in a way to enable **appropriate medical check-ups** to occur.
  Rationale: My health is paramount and my health records are currently all over the place. I am not confident I understand my own medical needs.

<u>Example of enablers:</u>

- I will **put aside regular time** to ensure my health records are managed properly and **reviewed for action**.
  Rationale: Once I tidy my records up, unless I commit to a regular management routine, I will miss appointments and return to my old untidy ways.

Data Governance for Community Groups

<u>Examples of decisions / policies</u>:

- Member details are to be managed in a way to enable **reliable** and **easy contact** of **current** members.
  Rationale: Without members, there is no group. Our members are valuable to keep the group alive. They may be invited to events or meetings or be contacted to show small acts of kindness such as when they are in need of personal support – eg grief, medical trouble, moving house -. or to congratulate them on something special in their life (eg birthdays)

<u>Example of enablers:</u>

- The **secretary** is ultimately responsible for ensuring that the member details are managed in such a way. To help with such task, the secretary is able to call upon other office bearers.
  Rationale: Whilst only one person must be responsible and accountable, some of the activities required to maintain a high standard of membership details may need the involvement of others on an ad-hoc or regular basis.

Data Governance for Businessess

Examples of decisions / policies:

- The management of our **Parts** is paramount and must **mirror the supplier catalogue at all times.**
  Rationale: Our customers expect parts details to be available for discussion at all times even though they may not be in stock and may need to be ordered.
- Our **Customer** data is **central to all our activities** and is managed by one group only: **Sales**.
  Rationale: To ensure the quality of the Customer data, ownership has to be allocated to one group. The Sales group is such group as it usually has first contact with new clients and is therefore expected to gather most Customer data as a usual core activity.

Examples of enablers:

- Analysis, design, implementation, and testing of the system to be implemented to enable the supplier's parts catalogue to be exchanged electronically is allocated an initial budget as well as a budget for the ongoing update. Responsibility for such project lies with the **Parts** department.
  Rationale: The Parts department has the best business knowledge to drive this project which may or may not be fully carried out by the Parts team.
- The **Sales** group is allocated a further budget to ensure **data quality** and **sharing** of such customer data across the business. Such budget also covers the **training cost of staff across the organisation**.
  Rationale: Data quality processes are new processes which come at a further cost. Such processes include the design, implementation, and ongoing reporting and training of staff to ensure ongoing data quality.

Data Architecture

Data Architecture gets its input from the decisions / policies endorsed by the Data Governance group – although there is likely to be an exchange of information between the Data Governance and Data Architecture groups / activities to inform and shape such decisions.

Its outputs define how data is to be collected in broad terms, how it is retrieved, for reporting or update, how it is removed and/or archived including the auditing requirements, and who are the people or systems allowed to perform such activities.

Examples of decisions influencing the Data Architecture:

- The Sales manager is the designated owner of the Customer data.
- The Customer data is to be shared across all business activities.
- The Supplier Parts catalogue is the source of information for our parts.
- The following legislation is to be adhered to:
  - Taxation Office record keeping for business requirements (Source: https://www.ato.gov.au/Business/Record-keeping-for-business/)
  - Privacy Act 1988 so that proper controls are in place to limit access to sensitive or personal data. (Source: https://www.oaic.gov.au/privacy/the-privacy-act/ and https://business.gov.au/Risk-management/Cyber-security/How-to-protect-your-customers-information)
- The following standards are to be adhered to:
  - AS 4590.1:2017: Interchange of client information, Part 1: Data elements and interchange formats (Source: https://www.standards.org.au/standards-catalogue/sa-snz/communication/it-004/as--4590-dot-1-colon-2017)

Examples of output from the Data Architecture group:

- Documentation of the parts of the AS 4590 standard to be implemented.
- A common vocabulary of business terms so that client, and customer are understood to be the same; so that a 'dog' and 'dog-trailer' in a livestock trucking business is well understood to mean the same, and not to be confused with the 4 legged animal which may be referred to as 'working dog' and would have different maintenance regime as an asset.
- Documentation of the technical means to access the common data such as Customer data.

Data Architecture for Individuals

Example of decisions / policies:

- All the Health data is to be kept **electronically** and be **secure**.
  Rationale: Electronic records can be easily carried to a doctor's surgery and be easily shared or just shown and discussed. Such records have to remain secure from theft and destruction such as fire.
- All **paper** documents are to be scanned and **originals securely kept**.
  Rationale: Scanning and security measures stem from the decision of keeping all records electronically and secure.

Data Architecture for Community Groups

Example of decisions / policies:

- All the Member data is to be kept **electronically** and be **secure**.
  Rationale: Electronic records can be easily carried to a doctor's surgery and be easy shared or just shown and discussed. Such records have to remain secure from theft and destruction such as fire.
- A **clearly dated list** of such details is made available to all office bearers **when needed** – whether **electronically** or **paper-based**.
  Rationale: Our group caters for all skill levels and our office bearers are not expected to be computer-literate. The date is important to ensure the same version of details is in use.

Data Modelling and Design

Your daily activities involve the usage of data – whether you buy a new appliance and collect a warranty card; whether your group renews its public liability insurance; or whether your business issues a new Tax Invoice. For each activity you must decide what happens to that data whether thrown in the bin to later realising it was important! or classing it in a paper folder or scanning it to store it electronically with or without associated details which you may want to further interrogate over time.

For example, you just renewed an insurance policy and you know it is important to keep the record in case of a future claim or to show proof to some interested party. So you decide to scan it and store it on your main electronic device in an area dedicated to this financial year. You also decide to keep a few specific details such as the purpose of the policy, policy number, due date, amount, paid date in a spreadsheet holding all your assets so that you can interrogate it at any time and work out when the policy is due and how much you paid last so you can budget for it.

The process you follow when you work out which specific details you keep (here: purpose of the policy, policy number, due date, amount, and paid date) and why you keep them is analytical. You analyse your data needs. You work out which activities you want to do such as budgetting and which details you need for this. This leads to defining the specific data items you wish to be able to electronically report on. This is the part of the data modelling activity known as **analysis of the data requirements**.

The **purpose** of the Data Modelling and Design activity is to **accurately define** the data you care to store electronically, including the applicable data quality rules, reporting rules, and other business rules. The activity may also include the **automated build** of the database – if a database is the target technology and a Computer-Aided System Engineering (CASE) tool is being used. (Refer to the Metadata section for managing data definitions.)

The examples below will also show how a data definition is linked to one or more process definitions. For each data item, there is a need to consider how this data item is **created**, **retrieved** for usage, **updated**, **deleted** from current records, and/or **archived** for later retrieval. And in all these data processes, one has to define **who** is allowed to carry out such activities.

Finally note that **defining data and their rules** goes hand in hand with **defining processes** the data will be involved in. Some processes are defined first and their data needs follow such as *creation of health provider details*. Some data needs create the needs for new processes such as **data quality** processes eg. *address verification*. (Refer to the Data Quality section.)

Data Modelling and Design for Individuals

Example of definition of the data needs:

For each of my health provider, I need to store the title and full name, the name of the provider, as well as the address of the practice, the telephone and fax numbers, and the provider number.

Rationale: Each time I need a referral from a specialist, my GP asks which health provider I wish to use and expects the title and full name, the address of the practice, and a telephone number. Occasionally I am asked for a fax number for the referral to be faxed to. The provider number is rarely used but I have had occasions where an imaging company would ask for the provider number to search in their database and ensure it is the appropriate health provider.

Data Storage and Operations

The Data Storage and Operations activity refers to the management of the **physical assets** on which the data is stored be they electronic devices but also software in use. Such assets have to be maintained, and secured, and sometimes require replacement due to the introduction of a new technology.

Data Storage and Operations for Individuals

Example of decisions / policies:

- The software to store the data is reliable and open source.
  Rationale: There is no budget for commercial software other than what is necessary.
- The data is stored on at least two devices, one acting as the main device – here a USB drive, and the second as the backup – here the laptop. Devices are to be stored in different locations.
  Rationale: Data is an asset which has to be secured against fire, theft, but also against physical failure of the device.

Data Security

Data Security deals with enforcing the appropriate access rights defined during the Data Modelling and Design phase. Not only it has to control **who** can access but also the **level** of access of each individual or system (indeed some systems also automatically access the data of other systems).

Legislation such as the Privacy Act defines how data can be used.

Data Security for Individuals

Example of decisions / policies:

- Comments based on my personal experience with the provider is NOT to be included in any report or document.
  Rationale: Such comments are based on my personal experience and are not seen as relevant to other consumers at this stage.
- Only I am able to create, update, delete, or archive the details of each health provider.
  Rationale: I am solely responsible for the quality of the health provider data.
- Health providers details as per the agreed report format are made available to the public.
  Rationale: If the publicly available details I gather can help others, I am happy to share them in the format I design and on the platform of my choice.

Data Integration and Interoperability

The scenarios in scope for the activities of Data Integration and Interoperability include:

- when an external system and an internal exchange data automatically
- when internal systems use the same source of data (for example, one set of Customer data shared across an entire organisation)
- when data is meant to be shared between two separate entities – even without automation - (for example, usingAustralian Post Postcode data in a Customer database)

In such situation rules have to be put in place to define the frequency of the exchange, the quality processes, the additional data which may be needed to support the quality of the exchange, the potential transformations the data has to go through when acquired etc. The activities enabling the data to go from system A to system B in a quality and repeatable fashion are the scope of this phase.

Data Integration and Interoperability for Individuals

Example of decisions / policies:

- The output destined to be shared with external consumers is to be produced as .PDF unless it is not possible.
  Rationale: One cannot assume that the software we use is available to the person trying to use the file. .PDF is a Portable Data Format available for free for all to use.

Documents and Content

This activity considers data found in electronic files (such as images, and various scanned documents such as signed membership forms, insurance policies, responses from Government or other bodies, ...) and paper records and how this data is integrated or not with the data which has been deemed important to be modelled in the Data Modelling and Design phase (also known as **structured data** as opposed to Documents and Content considered as **unstructured data**).

Documents and Content for Individuals

Example of decisions / policies:

- Each referral I get from my GP to a specialist is to be scanned and stored electronically in a way enabling easy retrieval.
  Rationale:By scanning the referral I have the ability to email it if required and I can also refer back to it for my own later access if need be once the referral has been handed over.
- Each scanned referral is named after the topic, the date, and the name of the provider, and stored in a folder called *Health Referrals.* Example*: mammogram_20201220_MarieCurie*
  Rationale: The name enables me to sort my referrals by their content first, then by a date relevance, and finally by the provider. Indeed my future usage is more likely to be focused on areas of my health and their timeline rather than on whom the health provider was – although I can do a visual search on this too.
- The folder structure to create for my health needs is:
  - folder: *MyHealth*
    Rationale: To separate documents relating to my health from other documents and make them easier to access.
  - subfolder: *Health Referrals*
    Rationale: To keep all my referrals in a central area. Easy to access. I may need to review this decision once I know the other health areas of interest such as reports, investigations.

Reference and Master Data

Reference and Master Data are data which is deemed **shared** across systems and need special management processes to maintain its **quality** across all such systems.

Typically Reference and Master Data are under the responsibility of one group, usually created by that group, and shared from that group.

The rules for updating such data are often subject to much discussion. Indeed on the one hand, the update could be under the sole control of the responsible group. However reliable and timely processes have to be put in place to communicate the update, get it quality-checked, and then ensure it is available for use by the 'sub-systems'. On the other hand, the management of such data could be de-centralised leaving the 'sub-systems' responsible for creating and quality-checking new data prior to communicating the new data to the source of 'truth'. However how do you deal with the same data being created by another sub-system? These are some of the issues commonly discussed.

Reference data is data shared across systems but data which remains mostly unchanged over time.

Examples of 'external' Reference data is Australia Post postcodes. However this data is coming from an external source and the processes to be implemented fall under the Integration and Inter-operability section.

An example of 'internal' Reference data is the list of Australian states. Such list is well known, very stable, and managed internally ie there is no need to get an update from an external source.

Master data is data shared across systems but changing as business processes take place. For example, a business dealing with many customers will create new customer accounts on a regular basis. Such data is to be shared across internal systems but is regularly changing – adding and updating customer details.

Reference and Master Data for Individuals

<u>Example of decisions / policies</u>:

- I consider the list of health providers I deal with as Master data.
  Rationale: The health providers are used in the reporting I carry out on my health results, health events, and even financial reports. I want to be able to build a clear picture on each Health Provider. This assumes a single quality record for each health provider.

Data Warehousing and Business Intelligence

Electronic data-centric systems are usually either **optimised** for data entry and update, or reporting on large sets of data.

For example, to access a specific customer record for update, one has to either be able to quote and search on a specific customer number or on a greater number of details as we cannot rely on a name to be unique without considering other details such as date of birth. *(Note that this may not be unique either but will be enough to keep this example readable.)* To optimise the search, the system may be indexed on the most common search items being the combination of full names and date of birth, say. So, similar to an index in a book, rather than going through each page, the search goes straight to the index arranged in order of full name followed by date of birth, and a reference to the unique customer number being sought after. The index may look like:

John Smith, 21 January 1900, customer number: 413
John Smith, 21 January 1910, customer number: 91
John Smithe, 21 January 1900, customer number: 180

When data is accumulated over the years, the system may be asked to report on all customers born in a certain range of years, eg Generation Y, and list the 5 most popular product types they bought. Such searches may focus on years only and product type such as iPods. So they are indexed differently and require retrieval of many more records and amalgamation of such records so that they can report, for example:

Generation Y five most popular product types are:

iPod - 1,000,000
iPad - 500,000
smart phone – 300,600
TV – 100,000
washing machine – 50,0000

When systems are expected to support day-to-day transaction as well as the analysis of an enormous amount of data such as the tens of thousands of transactions of livestock sold in Australian weekly, or the millions of incoming and outgoing goods dealt with by Australian Customs, it is common to see two systems being designed: the **operational** day-to-day system storing mostly current data, and the **analytical** system accumulating years of data. The latter is usually known as a **data warehouse.**

Business Intelligence is the term commonly used when deriving knowledge from accumulated sets of data; knowledge which enables the business, group, or individual to perform better or provide a better service.

Let's explore some other examples.

Data Warehousing and Business Intelligence for Individuals

Example of decisions / policies:

- The details of health appointments in terms of date and cost are kept for 10 years. This includes the amount reimbursed by the social security, the amount reimbursed by my private health insurance, and the out-of-pocket amount.
  Rationale: I am interested in the trend and the particulars of my health costs so that I can make better decision when it comes to choosing the appropriate private health insurance provider and plan.

Meta-data

Meta-data is often referred to as 'data defining data' which sounds neither clear nor useful without an example.

In the context of a public library, a book is a physical asset which gets borrowed. When an electronic system is put in place to borrow books, each book becomes the **data** of interest. Typically the entire book is NOT scanned and made available in that format to the public. Instead some details about that book are selected and presented to the public so to describe the book well enough to support the process of borrowing books. The descriptive **meta-data** typically presented to a reader for each physical book (the data) includes the *title, author, publisher,* and *ISBN.*

There needs to be a way to document what a title looks like eg a string of characters up to 500 characters, say. Some metadata such as ISBN is defined by a known format too.

An ISBN is an International Standard Book Number. ISBNs were 10 digits in length up to the end of December 2006, but since 1 January 2007 they now always consist of 13 digits. ISBNs are calculated using a specific mathematical formula and include a check digit to validate the number.

(Source: https://www.isbn-international.org/content/what-isbn)

When defining a system known as a database to store the details for each book, we first need to decide which metadata is needed for each book (here, *title, author, publisher,* and *ISBN* as part of the data needs analaysis of the data modelling and design activity) and then we need to precisely define the format and acceptable values and rules defining the quality of the content for each of the metadata. These details must be documented as they form part of the data quality rules.

In other words the rules around ISBN could be simplified to just be a 10 digit or a 13 digit item. If an entry of 14 digits was input by an operator, the system would have to reject that entry.

The definition of the metadata for an ISBN can be documented in a plain document. When a database is professionally designed and deployed, a tool is used to manage the metadata. For example, Oracle Designer was in the 1990s a data modelling tool capturing metadata to support the data modelling and design activities for Oracle databases. Nowadays metadata tools can be used to describe metadata management needs at the enterprise level. Such tools are usually complex and all have their advantages and their drawbacks. Some are better suited to sharing definitions across an enterprise than others.

It is not expected that individuals, nor community groups would be literate in using metadata management tools. This said, basic database tools are metadata management tools guiding at least the physical design of the database; the mere input of column definitions is a metadata activity.

Meta-data for Individuals

Example of decisions / policies:

- Each Australian address is made up of four components being the address line/s, suburb, state, postcode.
  Rationale: There is no need to cater for overseas addresses, and there is no need to further split the address details as our system does not support geo-coding or mapping.

- Each Australian state is stored as a 2 to 3 letter acronym.
  Rationale: Australian states are defined by a known list of acronyms of up to 3 characters.

Data Quality

In the DAMA framework, a separate activity is dedicated for Data Quality and in my experience, this is not luxury. There is no point collecting data to support decision-making if the data is erroneous.

- as an individual you would systematically miss your health appointments,
- as a community group, you would not be certain of your membership details,
- as a business, you would be ordering the wrong parts and lose customers and monies over it.

Data Quality defines how quality is measured for each collected data item. The activity also defines the processes which will ensure that such quality is in place, and if not, how to address the issues. So there is an initial definition component which may be refined over time and an ongoing commitment to check the quality level of the data.

The processes to verify quality can be as simple as asking the customer to verify their own details and confirm their validity. It may also involve enforcing some quality rules around the data entry of addresses to ensure such addresses are valid and enable services to be delivered such courier or postal services.

Data Quality for Individuals

Example of decisions / policies:

[DQ-HP-ADDR] The address of each Health Provider is visually checked against an official source eg website, letterhead... before inputting in the electronic system.

Rationale: I will not attempt to use a system which automatically checks the validity of the addresses at time of electronic input but I will make sure I have a correct address by checking against a ***reputable*** *source. It will save me time when I visit the practice next as I will be confident that the address is correct.This is based on the assumption and acceptance that a practice does communicate a change of address to its patient base, else a new process would need to take place being the checking of the address each time it needs to be used!*