Data Management
Principles and Data Lifecycle

Dr Emma Murphy

Week 3, February 2021

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH

TU DUBLIN

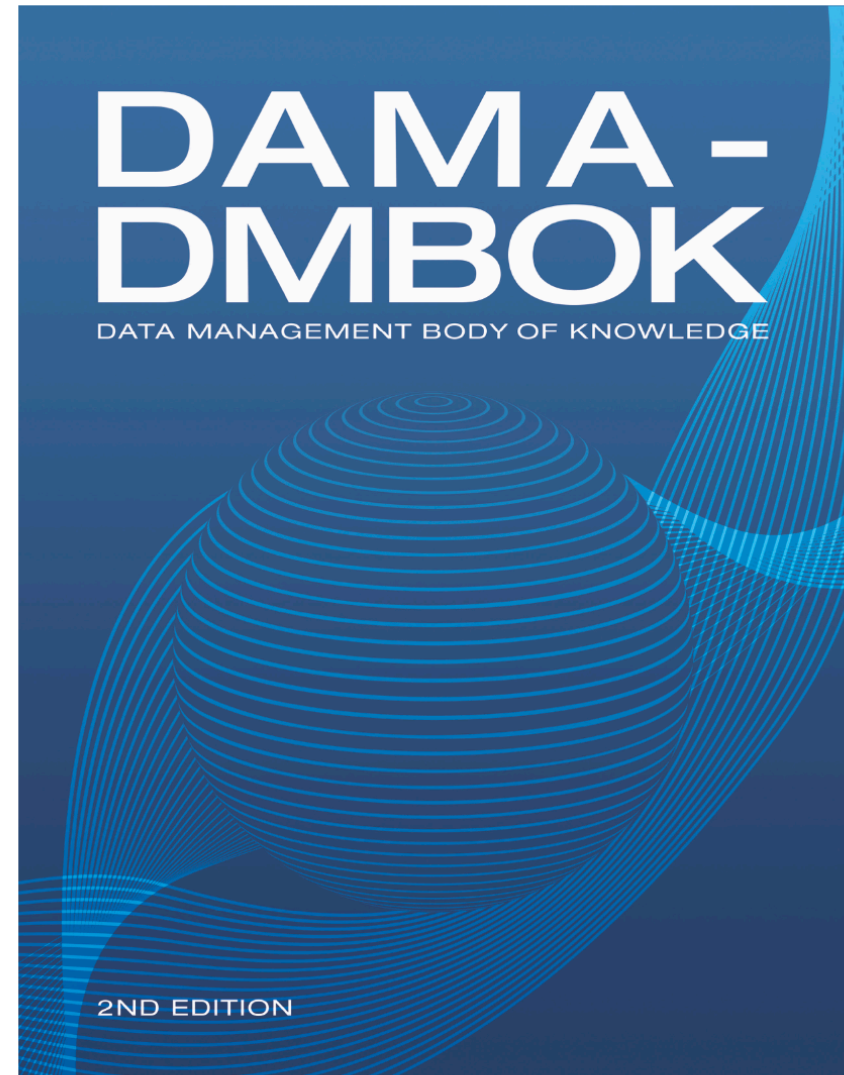TECHNOLOGICAL
UNIVERSITY DUBLIN

Recap

- Ethical Theories

- Data Ethics

- Practical ethical Assessment Framework

Overview

- Data Management Principles

- Data Management Lifecyle

- Digital Curation

- Data Management Strategies

DAMA DMBOK

- Data Management Association

- Data management body of knowledge

Data Management

- *Data Management* is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles.

(DAMA, 2017)

Data Management Professional

- A *Data Management Professional* is any person who works in any facet of data management

- Data management professionals fill numerous roles, from the highly technical (e.g., database administrators, network administrators, programmers) to strategic business (e.g., Data Stewards, Data Strategists, Chief Data Officers).

(DAMA, 2017)

Data Management Goals

- Understanding and supporting the information needs of the enterprise and its stakeholders

- Capturing, storing, protecting, and ensuring the integrity of data assets

- Ensuring the quality of data and information

- Ensuring the privacy and confidentiality of stakeholder data

- Preventing unauthorized or inappropriate access, manipulation, or use of data and information

- Ensuring data can be used effectively to add value to the enterprise

(DAMA, 2017)

Data as Representation

- Data is a means of representation- we need context for data to be meaningful.

- Context can be thought of as data's representational system; such a system includes a common vocabulary and a set of relationships between components.

- If we know the conventions of such a system, then we can interpret the data within it. These conventions are often documented in a specific kind of data referred to as Metadata.

(DAMA, 2017)

Data as Representation

- However, because people often make different choices about how to represent concepts, they create different ways of representing the same concepts.

- From these choices, data takes on different shapes. Think of the range of ways we have to represent calendar dates, a concept about which there is an agreed-to definition.

(DAMA, 2017)

Multiple representations

- Even within a single organization, there are often multiple ways of representing the same idea.

- Hence the need for Data Architecture, modeling, governance, and stewardship, and Metadata and Data Quality management, all of which help people understand and use data.

(DAMA, 2017)

Data Management Principles

- Data shares characteristics with other assets and therefore data management shares attributes with asset management. All management practices must balance strategic and operational needs.

- This balance can be struck by following principles that recognise key features of data management and guide data management practice:

(DAMA, 2017)

Data Management Principles

- **Data is an asset with unique properties**: Data is an asset, but it differs from other assets in important ways that influence how it is managed. The most obvious of these properties is that data is not consumed when it is used, as are financial and physical assets.

- **The value of data can and should be expressed in economic terms:** Calling data an asset implies that it has value. We should measure both the costs of low-quality data and the benefits of high quality data.

(DAMA, 2017)

Data Management Principles

- **Managing data means managing the quality of data:** Ensuring that data is fit for purpose is a primary goal of data management.

- **It takes Metadata to manage data:** Managing any asset requires having data about that asset (number of employees, accounting codes, etc.). The data used to manage and use data is called *Metadata*.

(DAMA, 2017)

Data Management Principles

- **It takes planning to manage data**: Data is created in many places and is moved between places for use. To coordinate work and keep the end results aligned requires planning from an architectural and process perspective.

- **Data management is cross-functional; it requires a range of skills and expertise**: A single team cannot manage all of an organization's data. Data management requires both technical and non-technical skills and the ability to collaborate.

(DAMA, 2017)

Data Management Principles

- Data management requires an enterprise perspective: Data management has local applications, but it must be applied across the enterprise to be as effective as possible. This is one reason why data management and data governance are intertwined.

- Data management must account for a range of perspectives: Data is fluid. Data management must constantly evolve to keep up with the ways data is created and used and the data consumers who use it.

(DAMA, 2017)

Data Management Principles

- Data management is lifecycle management: Data has a lifecycle and managing data requires managing its lifecycle.

- Different types of data have different lifecycle characteristics: Data management practices have to recognize these differences and be flexible enough to meet different kinds of data lifecycle requirements.

- Managing data includes managing the risks associated with data: In addition to being an asset, data also represents risk to an organization. Data can be lost, stolen, or misused. Data-related risks must be managed as part of the data lifecycle.

(DAMA, 2017)

Data Management Principles

- Data management requirements must drive Information Technology decisions: Data and data management are deeply intertwined with information technology and information technology management. Effective data management requires leadership commitment: Data management involves a complex set of processes that, to be effective, require coordination, collaboration, and commitment. Getting there requires not only management skills, but also the vision and purpose that come from committed leadership.

Data as an asset

- An *asset* is an economic resource, that can be owned or controlled, and that holds or produces value.

- Assets can be converted to money.

- Data is widely recognized as an enterprise asset, though understanding of what it means to manage data as an asset is still evolving.

Data is dynamic

- Data is dynamic and can be used for multiple purposes.

- The same data can even be used by multiple people at the same time – something that is impossible with physical or financial assets.

- Many uses of data beget more data. Most organizations must manage increasing volumes of data and the relation between data sets.

(DAMA, 2017)

Data is different to other assets

- Physical assets can be pointed to, touched, and moved around. They can be in only one place at a time.

- Financial assets must be accounted for on a balance sheet. However, data is different.

- Data is not tangible. Yet it is durable; it does not wear out, though the value of data often changes as it ages.

- Data is easy to copy and transport. But it is not easy to reproduce if it is lost or destroyed. Because it is not consumed when used, it can even be stolen without being gone

(DAMA, 2017)

Data valuation

- These differences make it challenging to put a monetary value on data. Without this monetary value, it is difficult to measure how data contributes to organizational success.

- These differences also raise other issues that affect data management, such as defining data ownership, inventorying how much data an organization has, protecting against the misuse of data, managing risk associated with data redundancy, and defining and enforcing standards for Data Quality.

(DAMA, 2017)

Data Valuation

- Despite the challenges with measuring the value of data, most people recognize that data, indeed, has value. An organization's data is unique to itself. Were organizationally unique data (such as customer lists, product inventories, or claim history) to be lost or destroyed, replacing it would be impossible or extremely costly.

- Data is also the means by which an organization knows itself – it is a meta-asset that describes other assets. As such, it provides the foundation for organizational insight.

(DAMA, 2017)

Data Valuation

- *Value* is the difference between the cost of a thing and the benefit derived from that thing.

- It is the difference between what the stock cost when it was purchased and what it was sold for.

- But for data, these calculations are more complicated, because neither the costs nor the benefits of data are standardized.

(DAMA, 2017)

Data Valuation

- Begin by articulating general cost and benefit categories that can be applied consistently within an organization. Sample categories include:
  - Cost of obtaining and storing data
  - Cost of replacing data if it were lost
  - Impact to the organization if data were missing
  - Cost of risk mitigation and potential cost of risks associated with data
  - Cost of improving data
  - Benefits of higher quality data
  - What competitors would pay for data
  - What the data could be sold for
  - Expected revenue from innovative uses of data

(DAMA, 2017)

Data Quality

- Ensuring that data is of high quality is central to data management. Organizations manage their data because they want to use it. If they cannot rely on it to meet business needs, then the effort to collect, store, secure, and enable access to it is wasted. To ensure data meets business needs, they must work with data consumers to define these needs, including characteristics that make data of high quality.

(DAMA, 2017)

Data Quality Issues

- What is the Data Quality Problem?

- Data is often:

- Inconsistent

- Inaccurate

- Unreliable

- Essentially, data doesn't mean what we think it does!

Data Quality Costs

- Scrap and rework

- Work-arounds and hidden correction processes

- Organizational inefficiencies or low productivity

- Organizational conflict

- Low job satisfaction

- Customer dissatisfaction

- Opportunity costs, including inability to innovate

- Compliance costs or fines

- Reputational costs

(DAMA, 2017)

Data Asset Management

- Data asset management ensures that all of an organization's data is treated as corporate assets with tangible value.
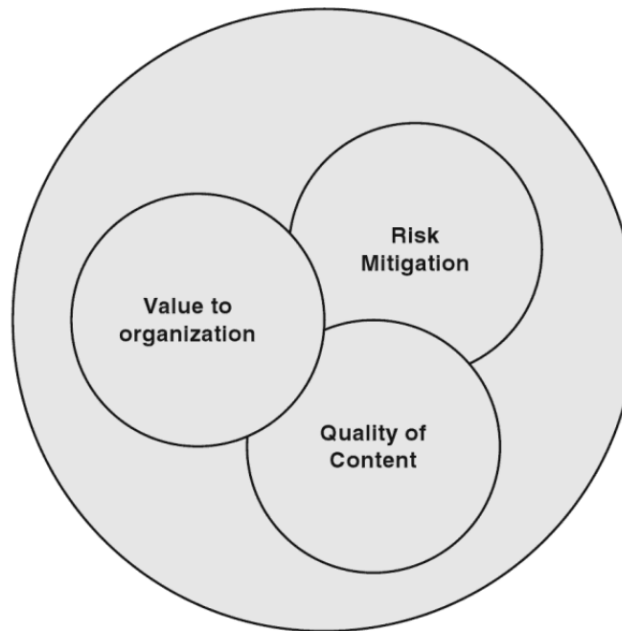
(Sarkar, 2015)

Data Lifecycle

- Like other assets, data has a lifecycle. To effectively manage data assets, organizations need to understand and plan for the data lifecycle.

- Well-managed data is managed strategically, with a vision of how the organization will use its data.

- A strategic organization will define not only its data content requirements, but also its data management requirements.
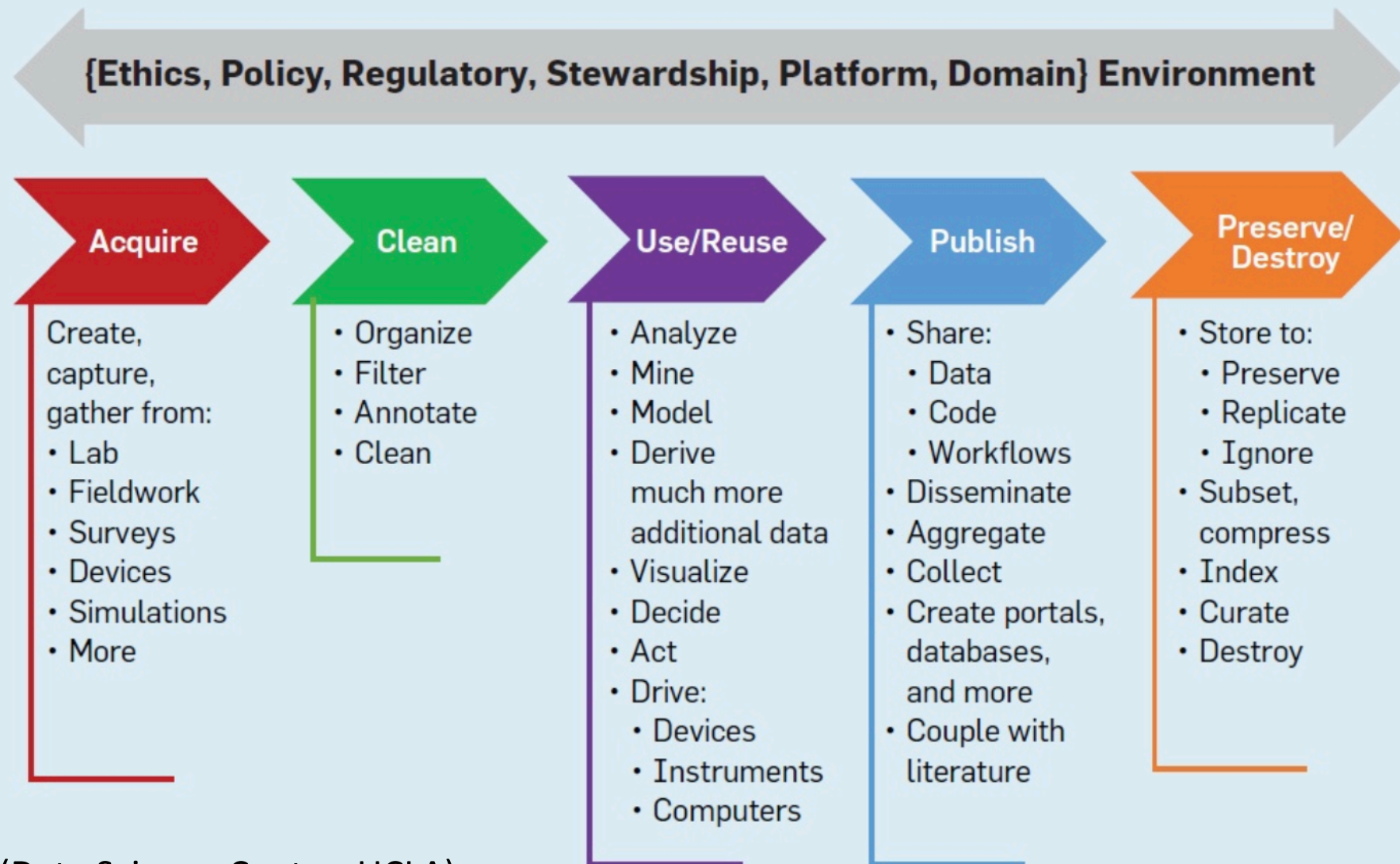
(DAMA, 2017)

Data Lifecycle

- The data lifecycle is based on the product lifecycle. It should not be confused with the systems development lifecycle.

- It includes processes that create or obtain data, those that move, transform, and store it and enable it to be maintained and shared, and those that use or or apply it as well as dispose of it.

(DAMA, 2017)

Data Lifecycle

- As data is used or enhanced, new data is often created, so the lifecycle has internal iterations that are not shown on the diagram.

- Data is rarely static.

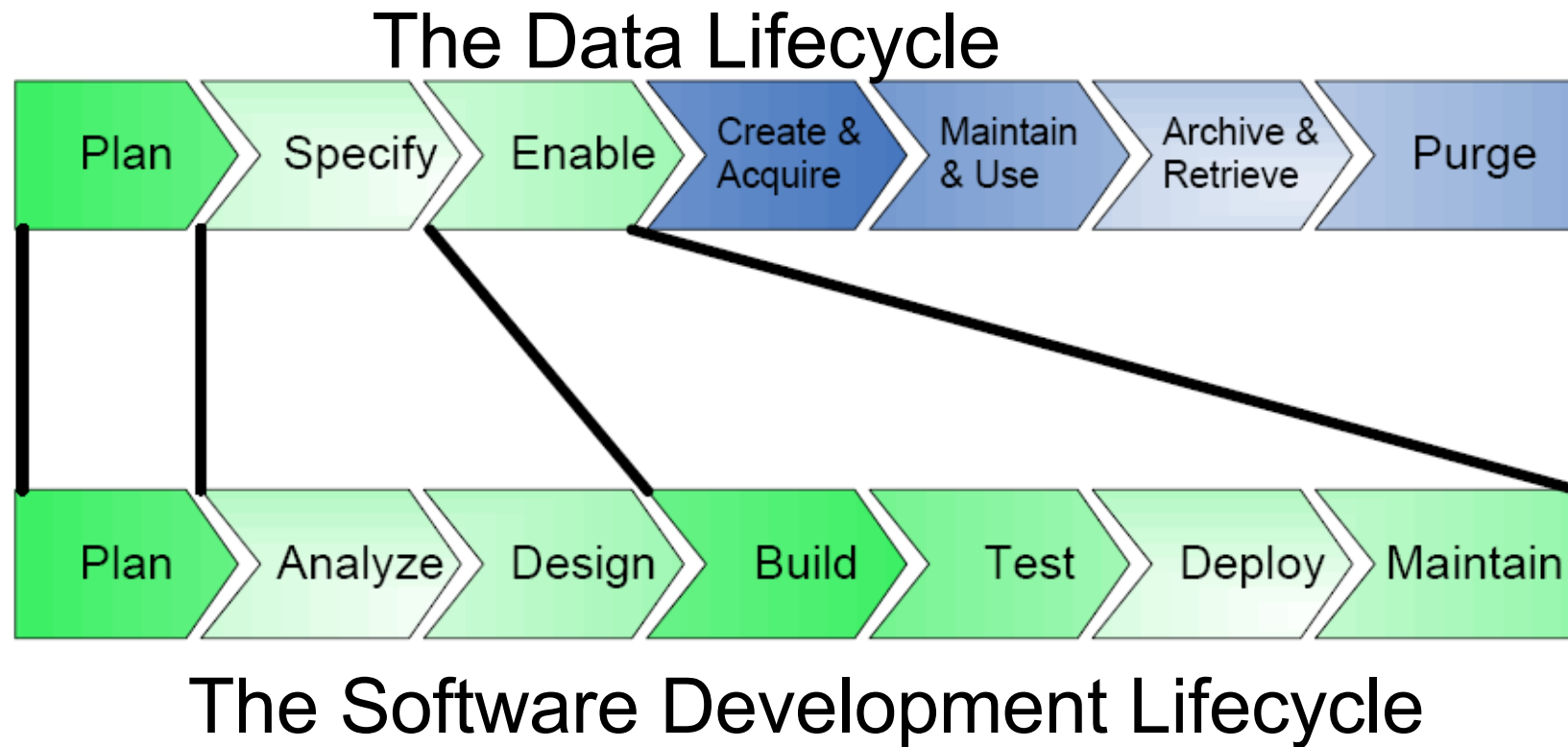- Managing data involves a set of interconnected processes aligned with the data lifecycle.

(DAMA, 2017)

Data Lifecycle

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire
Create, capture, gather from:
- Lab
- Fieldwork
- Surveys
- Devices
- Simulations
- More

Clean
- Organize
- Filter
- Annotate
- Clean

Use/Reuse
- Analyze
- Mine
- Model
- Derive much more additional data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

Publish
- Share:
  - Data
  - Code
  - Workflows
- Disseminate
- Aggregate
- Collect
- Create portals, databases, and more
- Couple with literature

Preserve/Destroy
- Store to:
  - Preserve
  - Replicate
  - Ignore
- Subset, compress
- Index
- Curate
- Destroy

(Data Science Centre, UCLA)

Data vs Software Lifecycle

The Data Lifecycle

| Plan | Specify | Enable | Create & Acquire | Maintain & Use | Archive & Retrieve | Purge |

| Plan | Analyze | Design | Build | Test | Deploy | Maintain |

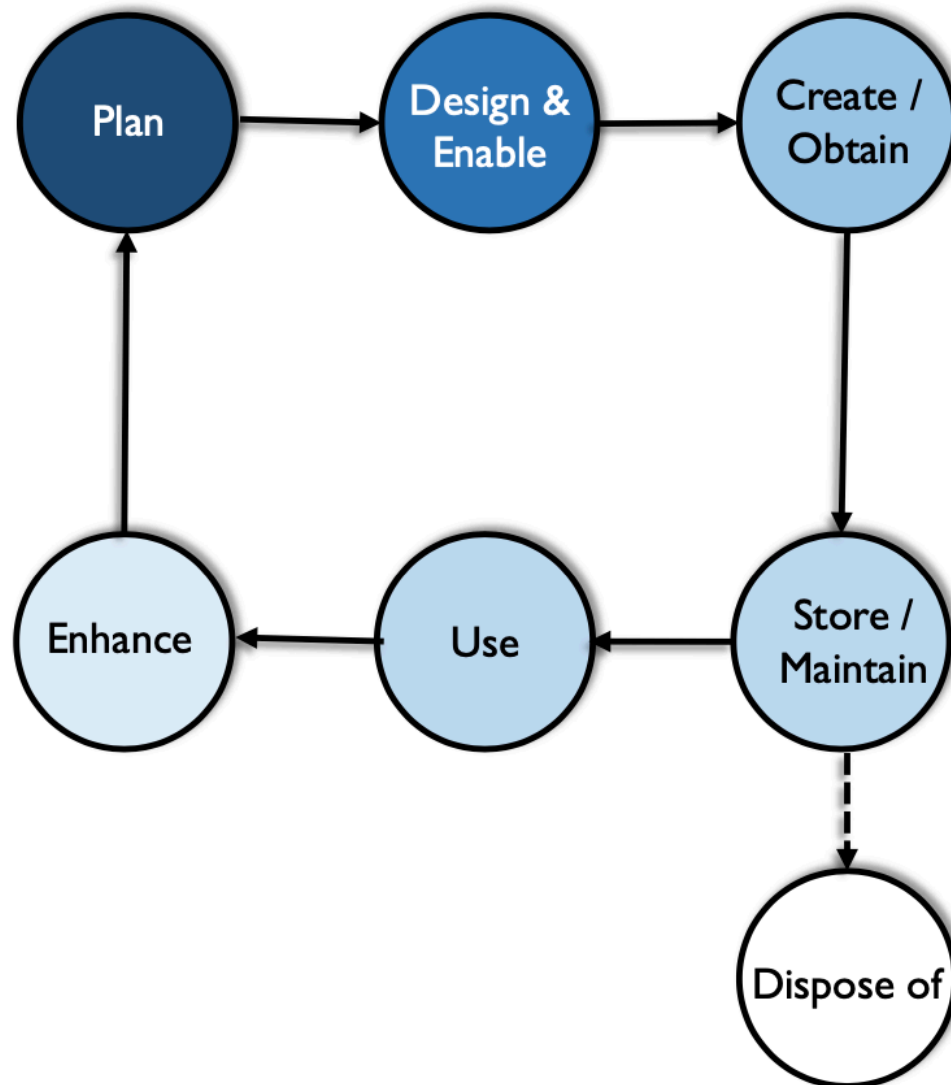The Software Development Lifecycle

(DAMA, 2009)

Software Lifecycles

Data Lifecycle

Data Lifecycle Key Activities
(DAMA, 2017)

Activity

- You are planning a research study that aims to investigate whether daily activity levels (measured using step count from an activity tracker) can predict rehospitalization of people who have suffered a major cardiac event.

- You will track the daily step count of 120 adults over 6 months. You will conduct monthly interviews with participants to elicit details of any exacerbations or hospitalizations.

- Can you create a data lifecycle diagram or visualization for the data in the proposed study?

Digital Curation

- For long term use, preserving the data and protecting it from loss or obsolescence, particularly if it is important, irreplaceable information, is essential in order to sustain the data assets.   "Digital curation is an ongoing process not a one-off action. It is a chain of activities only as strong as its weakest link". Abbot (2010)

- Part of digital curation is discovering and re using data and ensuring the information is in a readable format for future use and to ensure the data is a trust worthy reliable source.

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Digital Curation

- The Digital Curation Centre (DCC) - centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management.

- University of Edinburgh

- https://www.youtube.com/watch?v=6cuOdgvYRGM&feature=emb_logo

Research DM Lifecycle

http://www.dcc.ac.uk/resources/curation-lifecycle-model
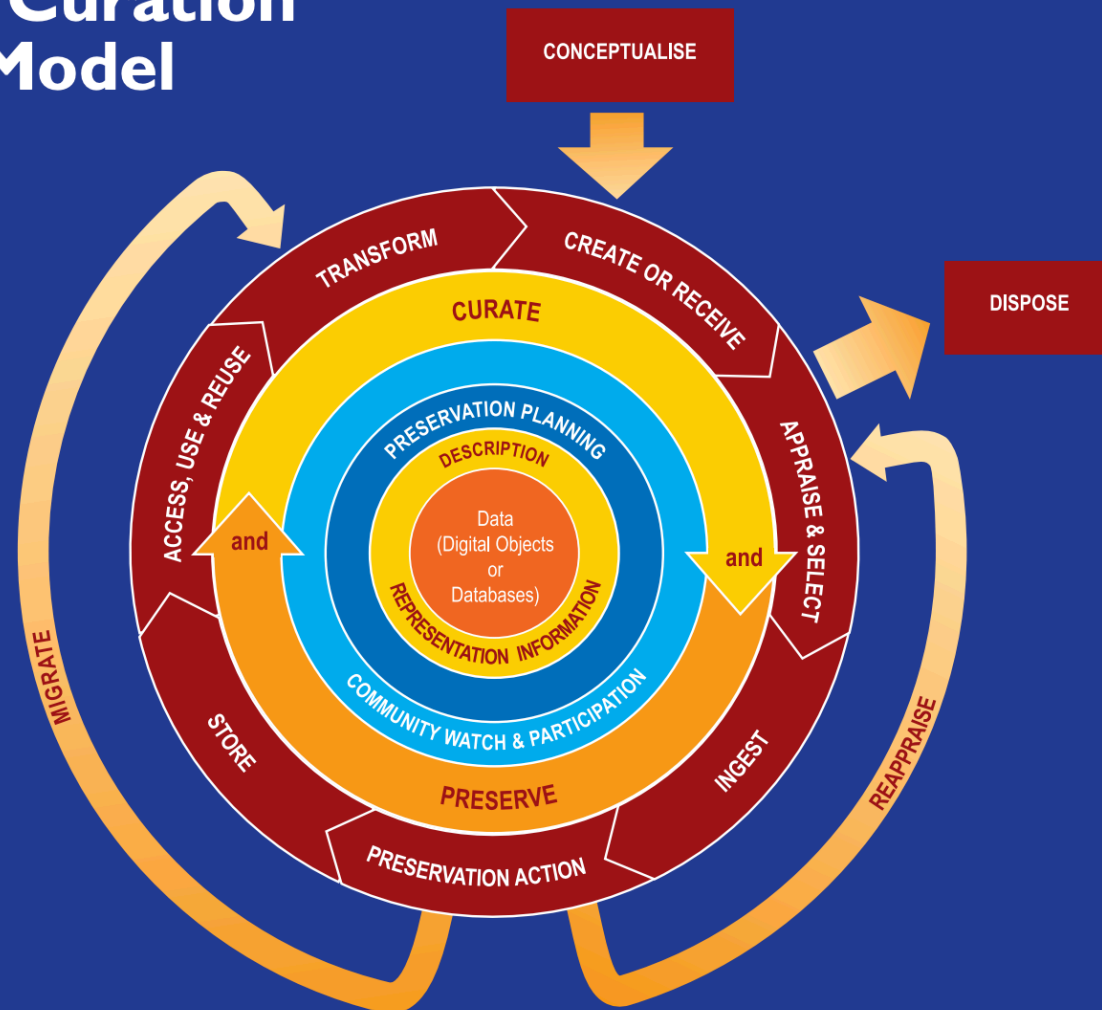
Data Curation – Lifecycle actions

- Description and Representation Information

  Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.

- Preservation Planning

  Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

- Community Watch and Participation

  Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

- Curate and Preserve

  Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Data Curation - Sequential Actions

- Conceptualise

  Create or Receive

  Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.

- Appraise and Select

  Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

- Ingest

  Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Data Curation - Sequential Actions

- **Preservation Action** - Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.

- **Store** - Store the data in a secure manner adhering to relevant standards.

- **Access, Use and Reuse** - Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.

- **Transform** - Create new data from the original, for example: by migration into a different format, or by creating a subset, by selection or query, to create newly derived results, perhaps for publication

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Data Curation - Occasional Actions

- **Dispose** - Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.

- **Reappraise** - Return data which fails validation procedures for further appraisal and re-selection.

- **Migrate** - Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Data lifecycle

- The focus of data management on the data lifecycle has several important implications:

- Creation and usage are the most critical points of the data lifecycle

- Data quality must be managed throughout the data lifecycle

- Metadata quality must be managed throughout the data lifecycle

- Data security must be managed through the data lifecycle

- Data management efforts must focus on the most critical data

(DAMA, 2017)

Data Disposal

- Effective data management requires clear policies and procedures, regarding retention and disposal of records.

- A retention and disposition policy will define the timeframes during which documents for operational, legal, fiscal or historical value must be maintained.

- This defines when inactive documents can be transferred to a secondary storage facility, such as off-site storage.

(DAMA, 2017)

Data disposal

- Non-value-added information should be removed from the organization's holdings and disposed of to avoid wasting physical and electronic space, as well as the cost associated with its maintenance.

- There is also risk associated with retaining records past their legally required timeframes. This information remains discoverable for litigation.

(DAMA, 2017)

Data Management Strategies

- A strategy is a set of choices and decisions that together chart a high-level course of action to achieve high-level goals

- A data strategy should include business plans to use information to competitive advantage and support enterprise goals.

- Data strategy must come from an understanding of the data needs inherent in the business strategy: what data the organization needs, how it will get the data, how it will manage it and ensure its reliability over time, and how it will utilize it.

(DAMA, 2017)

Data Management Strategy

- Typically, a data strategy requires a supporting Data Management program strategy – a plan for maintaining and improving the quality of data, data integrity, access, and security while mitigating known and implied risks. The strategy must also address known challenges related to data management.

- In many organizations, the data management strategy is owned and maintained by the CDO (Chief Data Officer) and enacted through a data governance team, supported by a Data Governance Council. Often, the CDO will draft an initial data strategy and data management strategy even before a Data Governance Council is formed, in order to gain senior management's commitment to establishing data stewardship and governance.

Data Management Strategy Components

- The components of a data management strategy should include:

- A compelling vision for data management

- A summary business case for data management, with selected examples

- Guiding principles, values, and management perspectives

- The mission and long-term directional goals of data management

Data Management Strategy Components

- Proposed measures of data management success

- Short-term (12-24 months) Data Management program objectives that are SMART (specific, measurable, actionable, realistic, time-bound)

- Descriptions of data management roles and organizations, along with a summary of their responsibilities and decision rights

- Descriptions of Data Management program components and initiatives

- A prioritized program of work with scope boundaries

- A draft implementation roadmap with projects and action items

(DAMA, 2017)

THE LEADER'S DATA MANIFESTO

A CALL FOR ACTION FOR MANAGING DATA ASSETS

Leader's Data Manifesto

- Data management practitioners met to discuss the state of managing data assets; what they found:

- The evidence of data's value is strong.

- The best practices are known.

- There are small-scale successes, but few that offer deep, lasting company-wide change.

Leader's Data Manifesto

- What they questioned:

- Why is adoption so slow?

- What are the root causes?

- And what needs to change?

Leader's Data Manifesto

- https://dataleaders.org/wp-content/uploads/2017/04/dataleaders-edw2017-sig-preso-manifesto-final-website.pdf

Activity

- Think about how you could map the data lifecycle of the product or service that you are researching for your assignment.

References

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

Sarkar, P. (2015) *Data as a Service: A Framework for Providing Reusable Enterprise Data Services*. John Wiley & Sons.