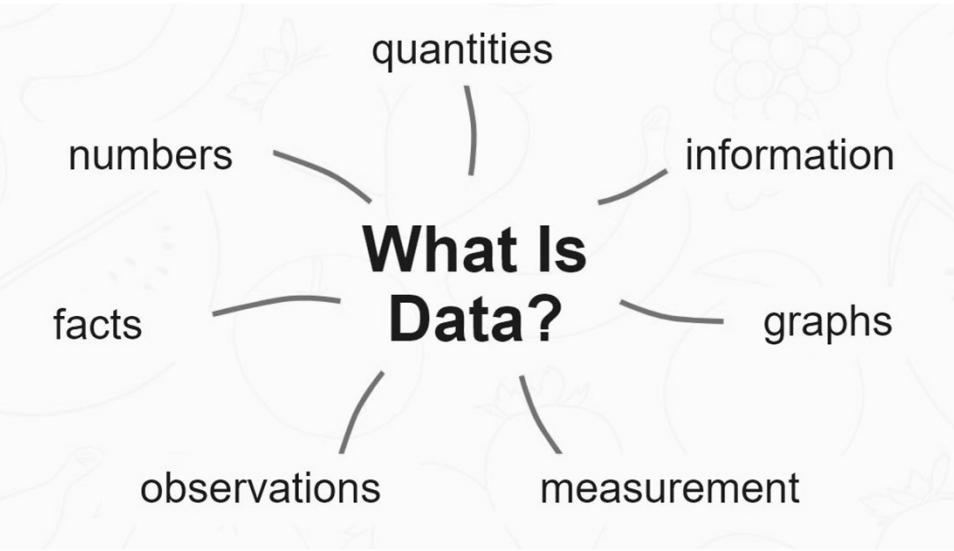
Data Governance

Vũ Tuyết Trinh

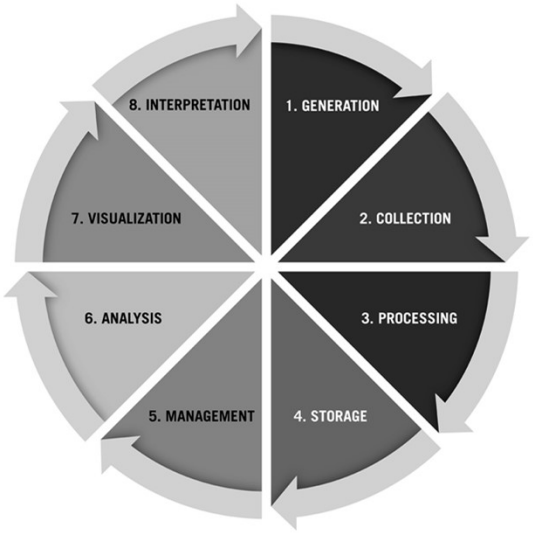
1



Source: https://images.twinkl.co.uk/tr/raw/upload/u/ux/what-is-data_ver_1.jpg

2

Data Life Cycle



3

Data Management (DM)

- managing data volume, variety, velocity, and veracity.
- Need of a scalability, flexibility, and robustness
- Integrating data from multiple sources, formats, and domains requires standardization, harmonization, and transformation of data.
- Ensuring data availability, accessibility, and usability requires data cataloging, metadata management, and data stewardship.

4

What is Data Governance ?

- DG is defined as the exercise of **authority** and **control** (planning, monitoring, and enforcement) over the management of **data assets**.

[DAMA Data Management Body of Knowledge V2 (DMBOK2)]

- DG encompasses the people, processes, and technology required to create a **consistent** and proper handling of an **organization's data across the business enterprise**.

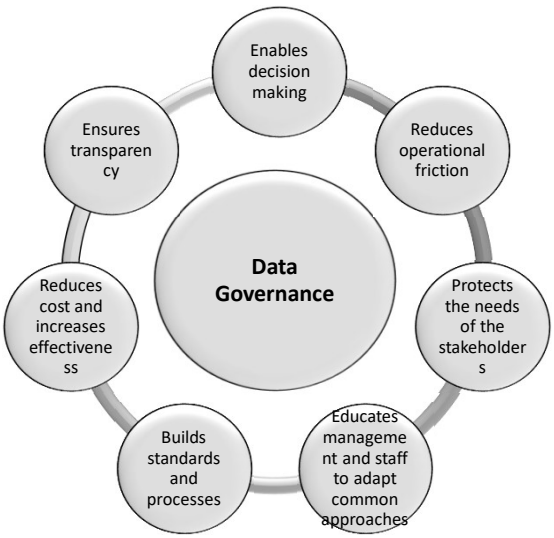
Wikipedia

- DG is the **orchestration** of people, **processes**, **policies** and technology to formally define, discover, assess, clean, integrate, and protect structured and unstructured **data assets** through their lifecycle to guarantee commonly understood, **trusted** and **secure** data throughout the enterprise

[Mike Ferguson, Intelligent Business Strategies]

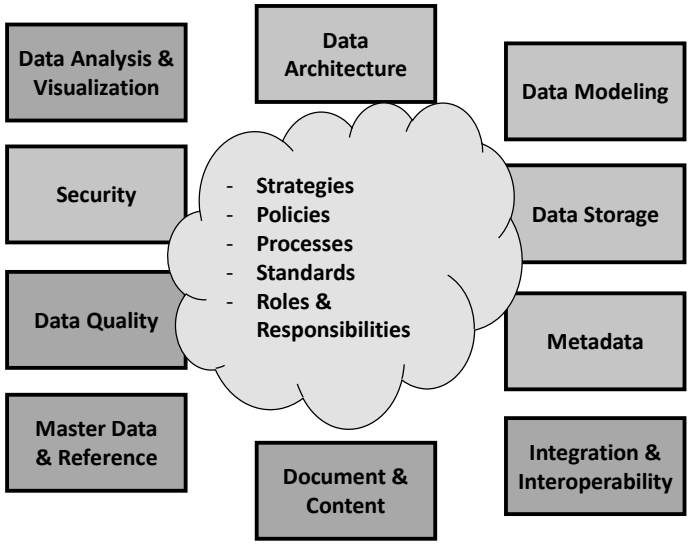
5

Goals



6

Data Governance (DG)



7

DM vs. DG

- How organizations use data
- All about execution, implementing business requirements: data engineer, architect, or DBA
- tools for data storage, processing, and exploration.
- How organization decide about using data
- business and IT teams's concerns: business managers, domain data owners, and other such business stakeholders
- rules and incorporate them for data assets across the organization.

8

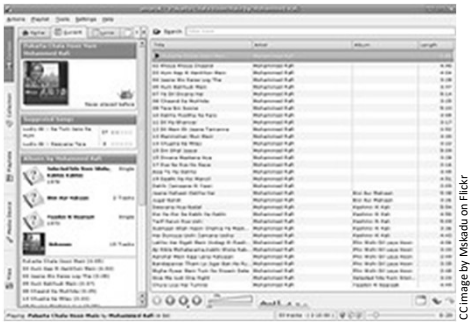
Metadata

Is data 'reporting'

- WHO created the data?
- WHAT is the content of the data?
- WHEN were the data created?
- WHERE is it geographically?
- HOW were the data developed?
- WHY were the data developed?

9

Metadata in Real Life



CC image by Mskadu on Flickr

Author(s)	Boullosa, Carmen.
Title(s)	They're cows, we're pigs / by Carmen Boullosa
Place	New York : Grove Press, 1997.
Physical Descr	viii, 180 p ; 22 cm.
Subject(s)	Pirates Caribbean Area Fiction.
Format	Fiction

Nutrition Facts

Serving Size 4 OZ. SERVING (112g)
Servings Per Container VARIED

Amount Per Serving

Calories 170	Calories from Fat 70
	% Daily Value*
Total Fat 8g	**12%**
Saturated Fat 3g	**15%**
Cholesterol 65mg	**22%**
Sodium 70mg	**3%**
Total Carbohydrate 0g	**0%**
Dietary Fiber 0g	**0%**
Sugars 0g	
Protein 23g	

Vitamin A 0% • Vitamin C 0%
Calcium 0% • Iron 15%

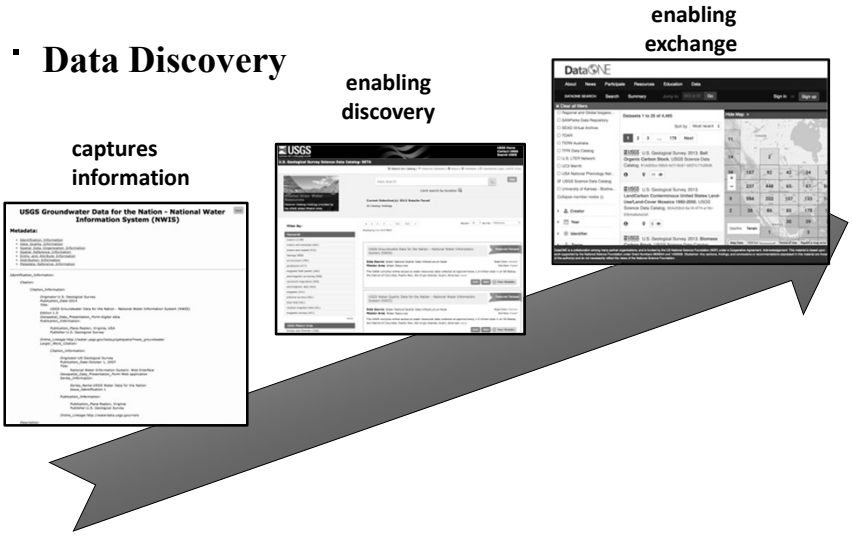
*Percent Daily Values are based on a 2,000 calorie diet.

CC image by USDAgov on Flickr

10

Metadata: What are they good for?

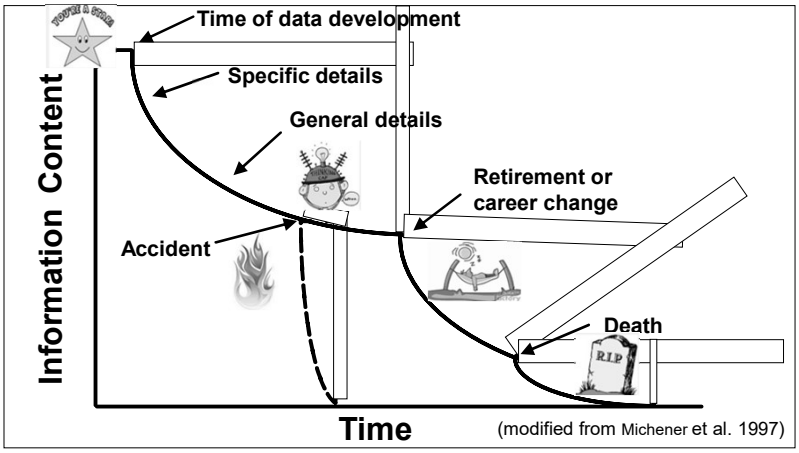
- **Data Discovery**



11

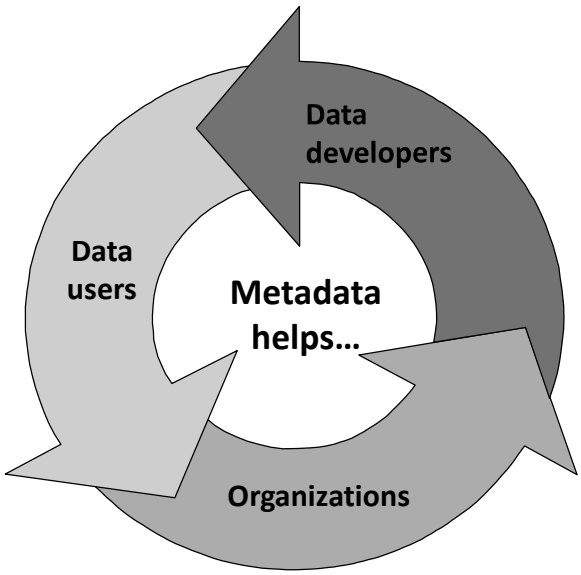
Metadata: Why are they important?

- **Scientific Understanding and Reuse**



12

The Value of Metadata



13

Metadata Standard

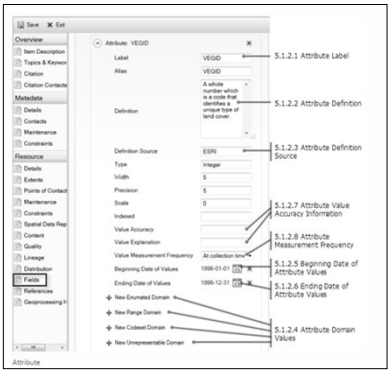
- A Standard provides a structure to describe data with:
  - Common terms to allow consistency between records
  - Common definitions for easier interpretation
  - Common language for ease of communication
  - Common structure to quickly locate information
- In search and retrieval, standards provide:
  - Documentation structure in a reliable and predictable format for computer interpretation
  - A uniform summary description of the dataset

14

What does a metadata standard include?

Components of metadata:

- A metadata standard is made up of defined elements, including the type of information the user should enter (e.g. text, numbers, date).
- Examples of elements include Title, Abstract, Keyword, Online Link



15

What Does a Metadata Record Look Like?

North American Breeding Bird Survey (BBS)

CC image by I like on Flickr

Identification Information:
Citation:
Citation Information:
Originator: Patuxent Wildlife Research Center, Biological Resources Division, U.S. Geological Survey (USGS)
Publication Date: 1997
Title:
North American Breeding Bird Survey (BBS)
Publication Information:
Publication Place: Laurel, MD
Publisher:
Patuxent Wildlife Research Center, Biological Resources Division, U.S. Geological Survey (USGS)
Other Citation Details:
This metadata file can be found at: ftp://cameron.cr.usgs.gov/pub/nbii_metadata/brdpwrc0004.txt (text format) and ftp://cameron.cr.usgs.gov/pub/nbii_metadata/brdpwrc0004.html (HTML format) and ftp://cameron.cr.usgs.gov/pub/nbii_metadata/brdpwrc0004.sgml (SGML format).
Description:
Abstract:
The North American Breeding Bird Survey (BBS), which is coordinated by the Biological Resources Division and Canadian Wildlife Service, is a primary source of population trend and distribution information for most species of North American birds. The BBS was initiated during 1966 by Chan Robbins and his associates at the Patuxent Wildlife Research Center to monitor the populations of all breeding bird species across the continental U.S., Canada, and Alaska. Approximately 2200 skilled observers participate in the survey each year. The BBS has accumulated 30 years of data on the abundance, distribution, and trends for more than 400 species of birds. These data are widely used by researchers, various federal and state agencies, non-governmental organizations, and the general public. Analyses of BBS data by PWRC statisticians have been instrumental in the development of innovative approaches for analyzing trends of wildlife populations.
Purpose:
In the 1960's, chlorinated hydrocarbon pesticides and similar poisons were widely used to control insect populations. Pesticide spraying not only killed insects but also killed birds, raising serious concerns over its effects on bird population trends. Unfortunately, no long-term regional or continental population data were available for most bird species, making it difficult for birders to demonstrate declines in bird populations. The Bird Breeding Survey has proven to be a valuable source of information on bird population trends. Robbins et al. (1986) provided the first continental relative abundance maps for various songbirds based on BBS data. When viewed at continental or regional scales, these maps provide a reasonably good indication of the relative abundance of species that are well sampled by the BBS. In addition, the BBS is a good source of information on temporal patterns in trends. Populations of permanent resident and short-distance migrant (birds wintering primarily in the

16

Metadata Standards - Examples

- **Dublin Core Element Set**
  - Emphasis on web resources, publications
  - http://dublincore.org/documents/dces/
- **FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
  - Emphasis on geospatial data
  - The Biological Data Profile (BDP) of the CSDGM is a profile to the CSDGM with an emphasis on biological data (and geospatial)
  - https://www.fgdc.gov/metadata/csdgm-standard

17

Metadata Standards - Examples

- **ISO 19115/19139 Geographic information – metadata**
  - Emphasis on geospatial data and services
  - https://www.fgdc.gov/metadata/iso-standards
- **Geography Markup Language (GML)**
  - Emphasis on geographic features (roads, highways, bridges)
  - http://www.opengeospatial.org/standards/gml

18

Metadata Standards - Examples

- **Ecological Metadata Language (EML)**
  - Focus on ecological data
  - http://knb.ecoinformatics.org/eml_metadata_guide.html
- **Darwin Core**
  - Emphasis on museum specimens
  - http://rs.tdwg.org/dwc/index.htm

19

Steps to Create Quality Metadata

1. Organize your information
   - Did you write a project abstract to obtain funding for your proposal? Re-use it in your metadata!
   - Did you use a lab notebook or other notes during the data development process that define measurements and other parameters?
   - Do you have the contact information for colleagues you worked with?
   - What about citations for other data sources you used in your project?

20

Steps to Create Quality Metadata

2. Write your metadata using a metadata tool
3. Review for accuracy and completeness
4. Have someone else read your record
5. Revise the record, based on comments from your reviewer
6. Review once more before you publish

21

Data Quality

- Quality assurance and quality control are strategies for
  - preventing errors from entering a dataset
  - ensuring data quality for entered data
  - monitoring, and maintaining data quality throughout the project
- Identify and enforce quality assurance and quality control measures throughout the Data Life Cycle
- QA/QC best practices
  - Before data collection
  - During data collection/entry
  - After data collection/entry

22

QA/QC During Data Entry

- Double entry
  - Data keyed in by two independent people
  - Check for agreement with computer verification
- Record a reading of the data and transcribe from the recording
- Use text-to-speech program to read data back

23

QA/QC During Data Entry

- Design data storage well
  - Minimize number of times items that must be entered repeatedly
  - Use consistent terminology
  - Atomize data: one cell per piece of information
- Document changes to data
  - Avoids duplicate error checking
  - Allows undo if necessary

24

QA/QC After Data Entry

- Make sure data line up in proper columns
- No missing, impossible, or anomalous values
- Perform statistical summaries

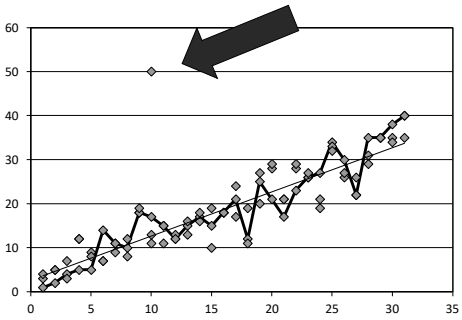
CC image by chesapeakeclimate on Flickr

25

QA/QC After Data Entry

Look for outliers

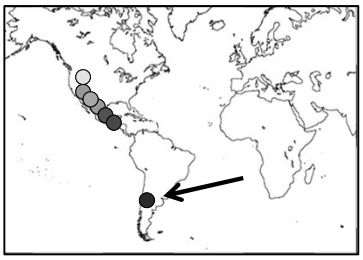
- extreme values for a variable given the statistical model being used
- identifying potential data contamination



26

QA/QC After Data Entry

- Methods to look for outliers
  - Graphical
    - Normal probability plots
    - Regression
    - Scatter plots
  - Maps
  - Subtract values from mean



27

Privacy and Security

- What we can collect and how
- How we share data, results and outcomes
- Reuse of human subject data
- Data storage and destruction
- IRB interpretations and review across institutions are not always consistent

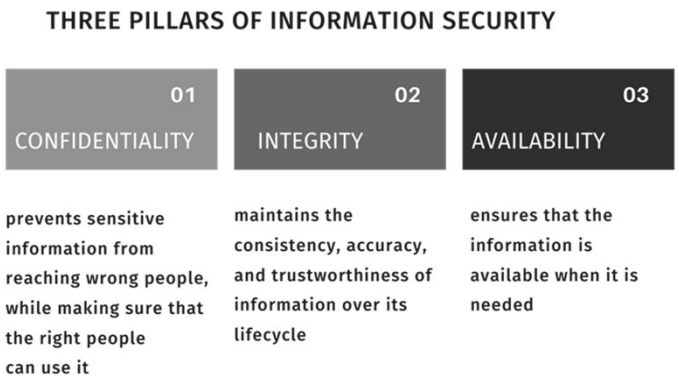
28

Privacy vs. Confidentiality

- Privacy
  - Protects access to individuals (or entities)
- Confidentiality
  - Protects access to information about individuals
  - Can be thought of as information privacy

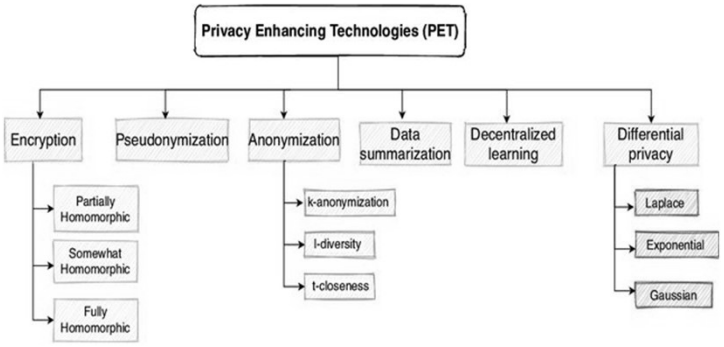
29

C-I-A triad



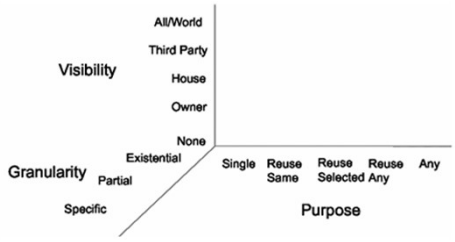
30

Privacy Technologies



31

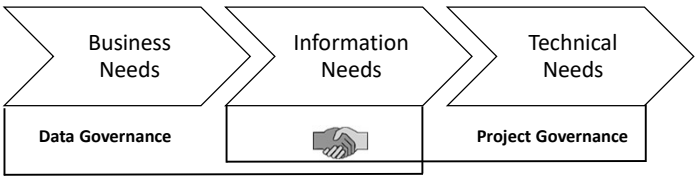
Data Privacy in a Data Repository



32

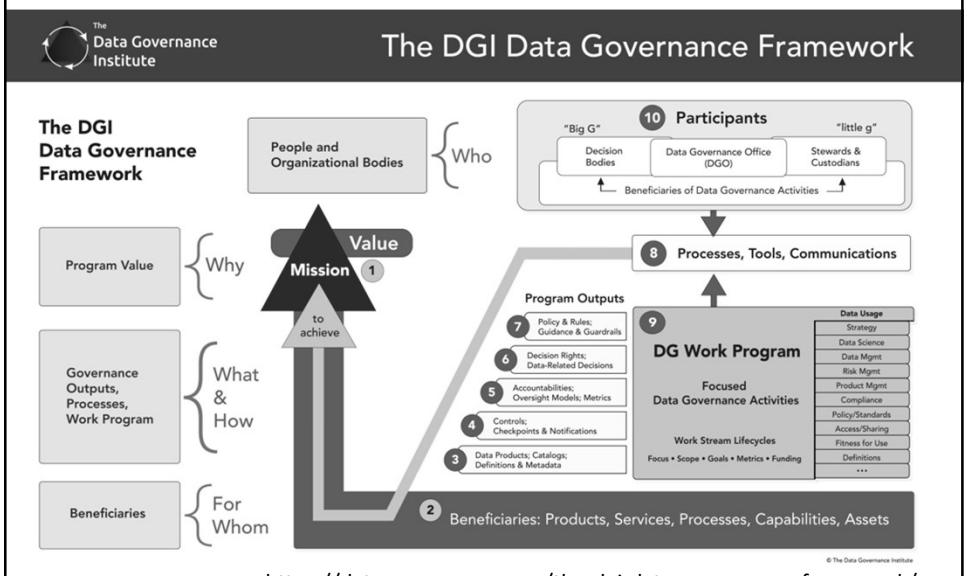
Data Governance Supports Strategic Business Goals

- Business needs drive information needs which drive technical needs (B.I.T.).



- Data Governance is primarily a business function and directly supports agency strategic goals.

33



https://datagovernance.com/the-dgi-data-governance-framework/

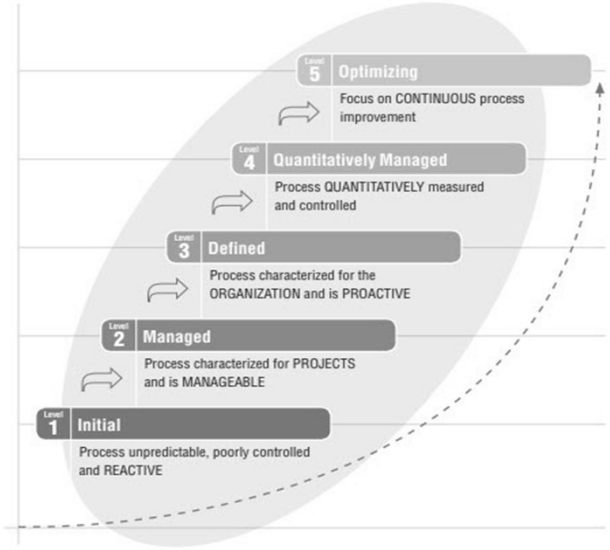
34

Data Governance Maturity

- Stage an organization has reached in the implementation and adaptation of Data governance initiatives
- Data governance maturity model is methodology to mesure organization Data governance initiatives

35

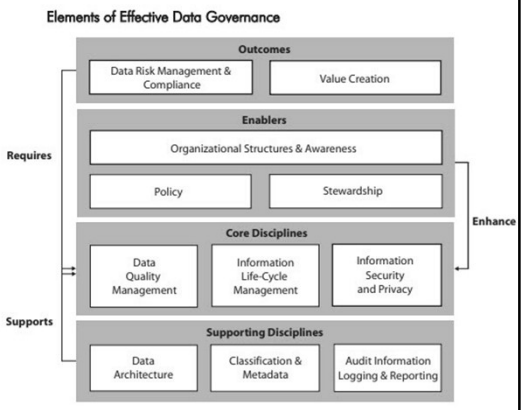
IBM model



36

11 data governance domains

- Data Risk Management & Compliance
- Value Creation
- Organizational Structures & Awaren
- Policy
- Stewardship
- Data Quality Management
- Information Lifecycle Management
- Information Security & Privacy
- Data Architecture
- Classification & Metadata
- Audit Information, Logging & Repo



37

References

- Plale, Beth & Kouper, Inna. (2017). The Centrality of Data: Data Lifecycle and Data Pipelines. https://online.hbs.edu/blog/post/data-life-cycle
- D. Edwards, in Ecological Data: Design, Management and Processing, WK Michener and JW Brunt, Eds. (Blackwell, New York, 2000), pp. 70-91.
- R. B. Cook, R. J. Olson, P. Kanciruk, L. A. Hook, Best practices for preparing ecological data sets to share and archive. Bull. Ecol. Soc. Amer. 82, 138-141 (2001).
- A. D. Chapman, "Principles of Data Quality:. Report for the Global Biodiversity Information Facility" (Global Biodiversity Information Facility, Copenhagen, 2004). Health information privacy. Accessed June 26, 2015 at http://www.hhs.gov/ocr/privacy.
- Protecting personal health information in research: understanding the HIPAA privacy rule. thttp://privacyruleandresearch.nih.gov/pr_02.asp.
- Constructing Access Permissions. http://libweb.uoregon.edu/datamanagement/sharingdata.html#three
- A taxonomy for privacy enhancing technologies (2015). https://www.sciencedirect.com/science/article/pii/S0167404815000668
- Barker, K., Askari, M., Banerjee, M., Ghazinour, K., Mackas, B., Majedi, M., ... & Williams, A. (2009, July). A data privacy taxonomy. In British National Conference on Databases (pp. 42-54). Springer, Berlin, Heidelberg.

38