DATA QUALITY

Objectives

- To explain the need for quality of data
- To describe how to ensure quality of data

Agenda

➢ **The Need for Quality of Data**

- Good Practices

Quality of Data is a Challenge

Factors that can undermine data quality:

- Lack of understanding about the effects of poor quality data on organisational success
- Bad or insufficient planning
- Isolated design of processes and systems
- Inconsistent technical development process
- Incomplete documentation and metadata
- Lack of standards and governance

Quality of Data is Important

Business Drivers

- Reduce risks and costs associated with poor quality data
- Improve organisational efficiency and productivity
- Protect and enhance organisation’s reputation

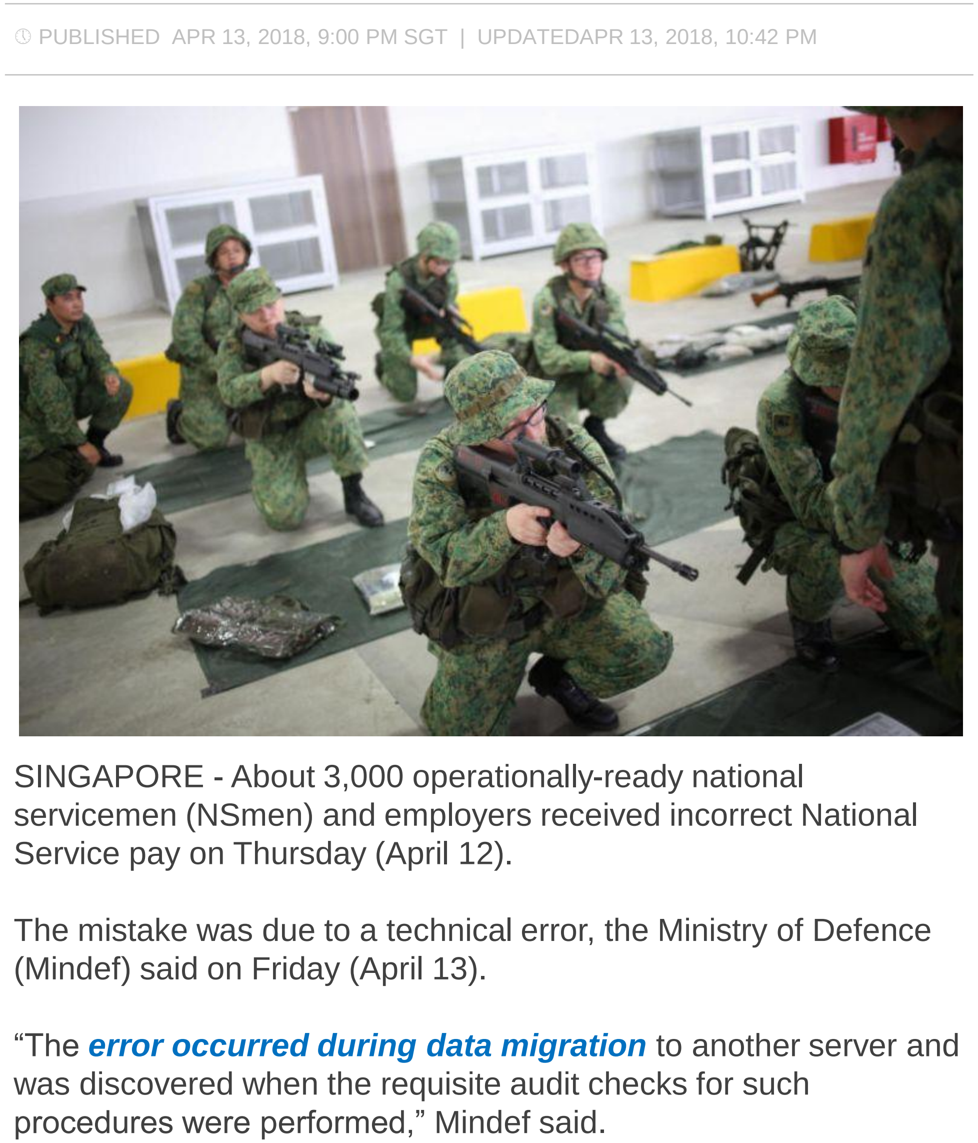
Undesirable Outcomes

- Erroneous invoicing
- Loss of customer
- Missed opportunities
- Costly integration
- Bad business decisions
- Exposure to fraud

RO THE STRAITS TIMES

THE BUSINESS TIMES

ursed

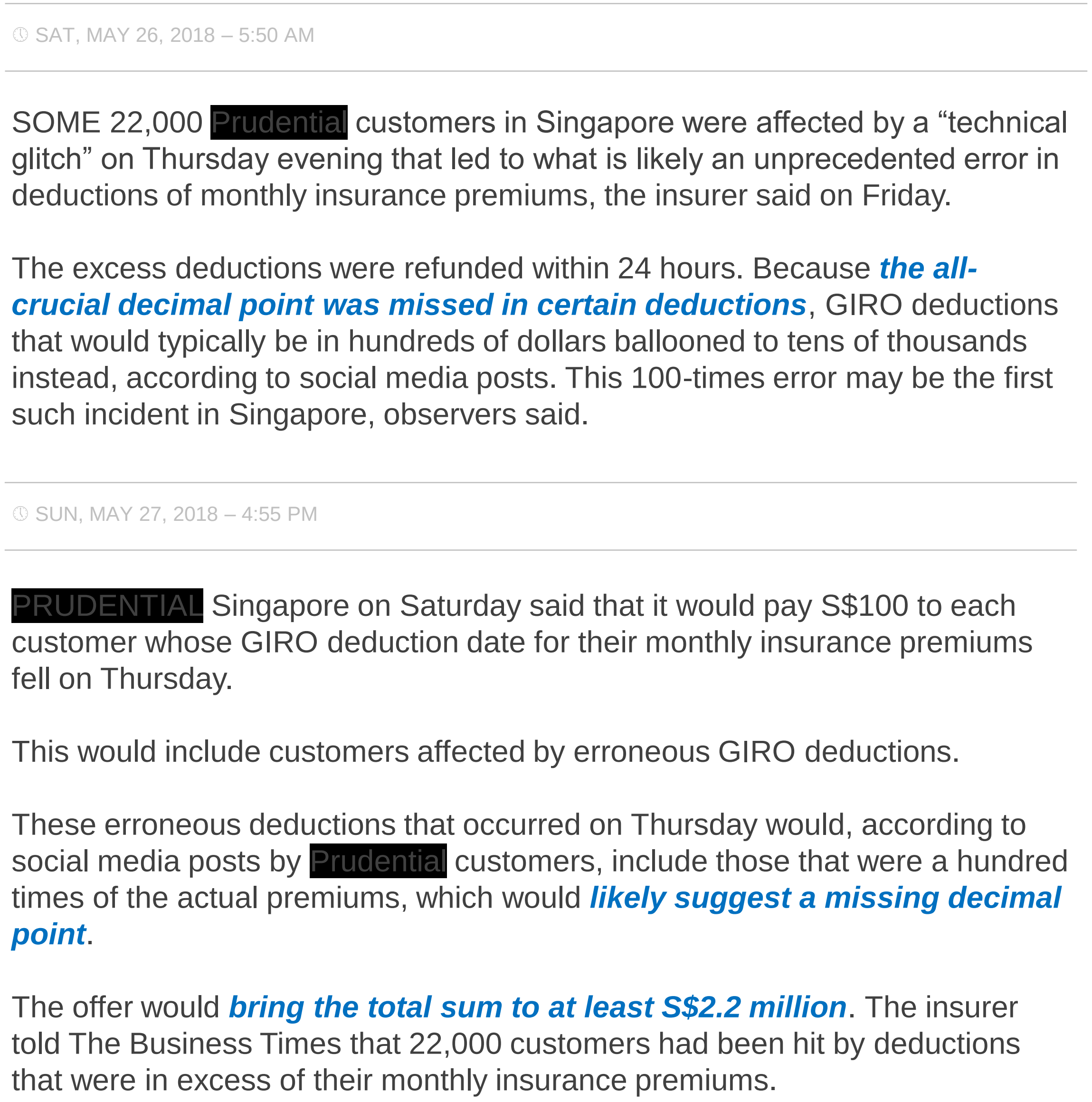


PUBLISHED APR 13, 2018, 9:00 PM SGT | UPDATEDAPR 13, 2018, 10:42 PM

SINGAPORE - About 3,000 operationally-ready national servicemen (NSmen) and employers received incorrect National Service pay on Thursday (April 12).

The mistake was due to a technical error, the Ministry of Defence (Mindef) said on Friday (April 13).

“The ***error occurred during data migration*** to another server and was discovered when the requisite audit checks for such procedures were performed,” Mindef said.



SAT, MAY 26, 2018 – 5:50 AM

SOME 22,000 Prudential customers in Singapore were affected by a “technical glitch” on Thursday evening that led to what is likely an unprecedented error in deductions of monthly insurance premiums, the insurer said on Friday.

The excess deductions were refunded within 24 hours. Because ***the all-crucial decimal point was missed in certain deductions***, GIRO deductions that would typically be in hundreds of dollars ballooned to tens of thousands instead, according to social media posts. This 100-times error may be the first such incident in Singapore, observers said.

SUN, MAY 27, 2018 – 4:55 PM

PRUDENTIAL Singapore on Saturday said that it would pay S$100 to each customer whose GIRO deduction date for their monthly insurance premiums fell on Thursday.

This would include customers affected by erroneous GIRO deductions.

These erroneous deductions that occurred on Thursday would, according to social media posts by Prudential customers, include those that were a hundred times of the actual premiums, which would ***likely suggest a missing decimal point***.

The offer would ***bring the total sum to at least S$2.2 million***. The insurer told The Business Times that 22,000 customers had been hit by deductions that were in excess of their monthly insurance premiums.

Responding to Regulation

MAS NOTICE 610

17 May 2018

NOTICE TO BANKS
BANKING ACT, CAP 19

SUBMISSION OF STATISTICS AND RETURNS

Introduction

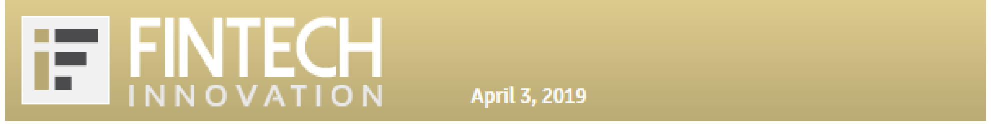
This Notice is issued pursuant to section 26(1) OF THE Banking Act (Cap. 19) ["the Act"]. It applies to all banks in Singapore.

…

3. A bank in Singapore shall submit the reporting forms to the Authority and ensure that each reporting form is approved by its chief executive or any person authorised by its chief executive to approve such reporting forms.

4. Subject to paragraphs 5 and 6,every bank shall lodge the reporting forms with the Authority as follows:

a) for the reporting forms which are to be submitted on a monthly basis…
b) for the reporting forms which are to be submitted on a quarterly or half-yearly basis…
c) for the reporting forms which are to be submitted on a yearly basis…


FINTECH INNOVATION April 3, 2019

Singapore banks sweating over MAS 610 compliance

By Fintech Innovation editors | 2018-06-30

Singapore banks point to their ability to successfully implement the Monetary Authority of Singapore (MAS) Notice 610 as their main concern for 2019.

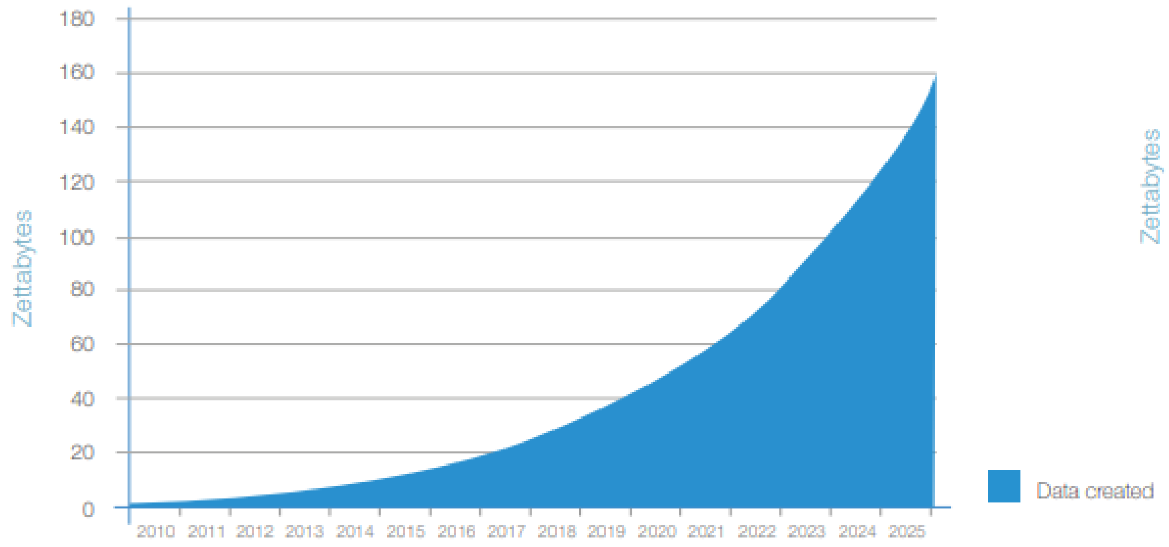
A survey of some 50 compliance, risk, finance and IT professionals at more than 25 banks in Singapore revealed that more than half of the banks surveyed identify the regulatory change as their top single concern for the year ahead according to a new survey by Wolters Kluwer's Finance, Risk & Reporting business.

The scope of new proposals in the Monetary Authority of Singapore's overhaul of the MAS 610 reporting regime for banks has taken many in the sector by surprise. The core set of returns that banks file to the Monetary Authority of Singapore are being revised to require information at a far more granular level beginning next year. In fact, ***the number of data elements that firms have to report will rise from about 4,000 to approximately 300,000***.

Twenty-seven percent of those surveyed pointed to having concerns around data analysis, gap identification and mapping, and ***data quality*** and remediation to meet the MAS 610 requirements, and 17% stated their top concern for the next year was simply keeping up with the pace of regulatory change.

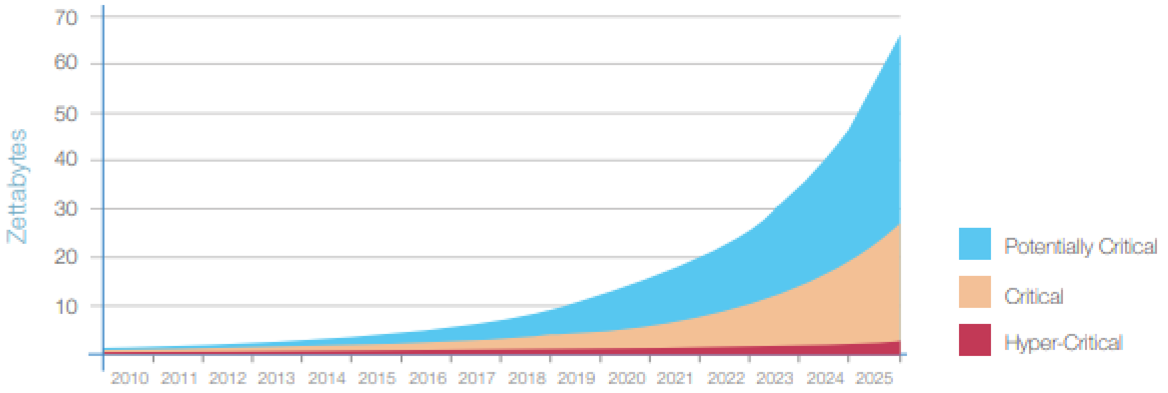
Compounded by Growth

Figure 2. Annual Size of the Global Datasphere



Source: Data Age 2025, IDC, sponsored by Seagate, Apr 2017

Figure 5. Data Criticality Over Time



Data Type
Potentially critical. Data that may be necessary for the continued, convenient operation of users' daily lives
Critical. Data known to be necessary for the expected continuity of users' daily lives.
Hypercritical. Data with direct and immediate impact on the health and well-being of users. (Examples include commercial air travel, medical applications, control systems, and telemetry. This category is heavy in metadata and data from embedded systems.)

Agenda

- The Need for Quality of Data
- ➢ **Good Practices**

Scope and Dimensions of Quality

Curate

- Behaviour in transactions
- Has a lifecycle
- Large set of elements
- Less volatility
- More complex
- More value
- Reusable
- Enterprise-designated critical data, e.g., for regulatory & financial reporting

Describe

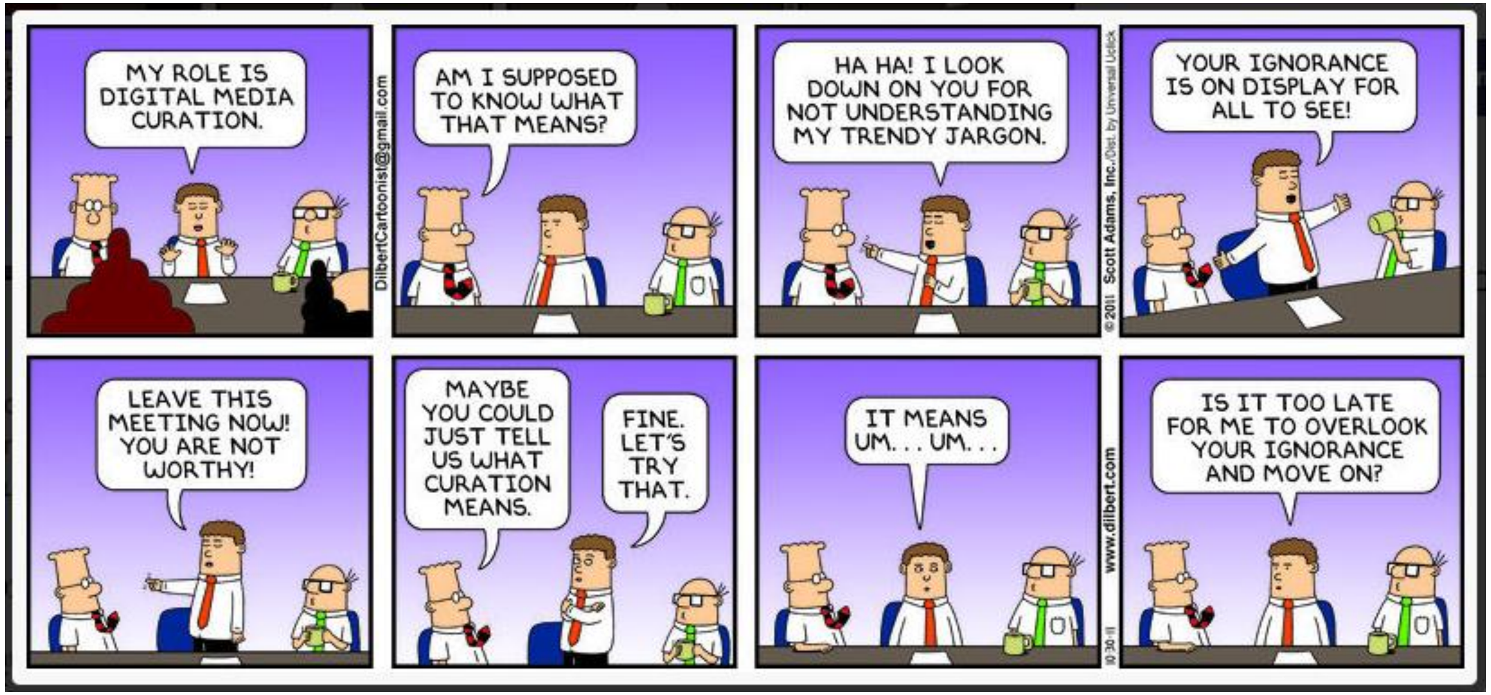
- Definition & description
- Business rules, transformation rules, calculations & derivations
- Data lineage
- Data standards
- Designation of system of record
- Value constraints
- Stakeholder information, e.g., Data Steward, Data Manager
- Security & privacy level

Profile & Clean[1]

- Completeness
- Accuracy
- Validity
- Consistency
- Integrity
- Uniqueness
- Timeliness

[1] if profiling leads to the identification of numerous issues, revisit curated data with stakeholders before proceeding

Data Curation


MY ROLE IS DIGITAL MEDIA CURATION.
AM I SUPPOSED TO KNOW WHAT THAT MEANS?
HA HA! I LOOK DOWN ON YOU FOR NOT UNDERSTANDING MY TRENDY JARGON.
YOUR IGNORANCE IS ON DISPLAY FOR ALL TO SEE!
LEAVE THIS MEETING NOW! YOU ARE NOT WORTHY!
MAYBE YOU COULD JUST TELL US WHAT CURATION MEANS.
FINE. LET'S TRY THAT.
IT MEANS UM. . . UM. . .
IS IT TOO LATE FOR ME TO OVERLOOK YOUR IGNORANCE AND MOVE ON?
DilbertCartoonist@gmail.com
©2011 Scott Adams, Inc./Dist. by Universal Uclick
www.dilbert.com
10-30-11

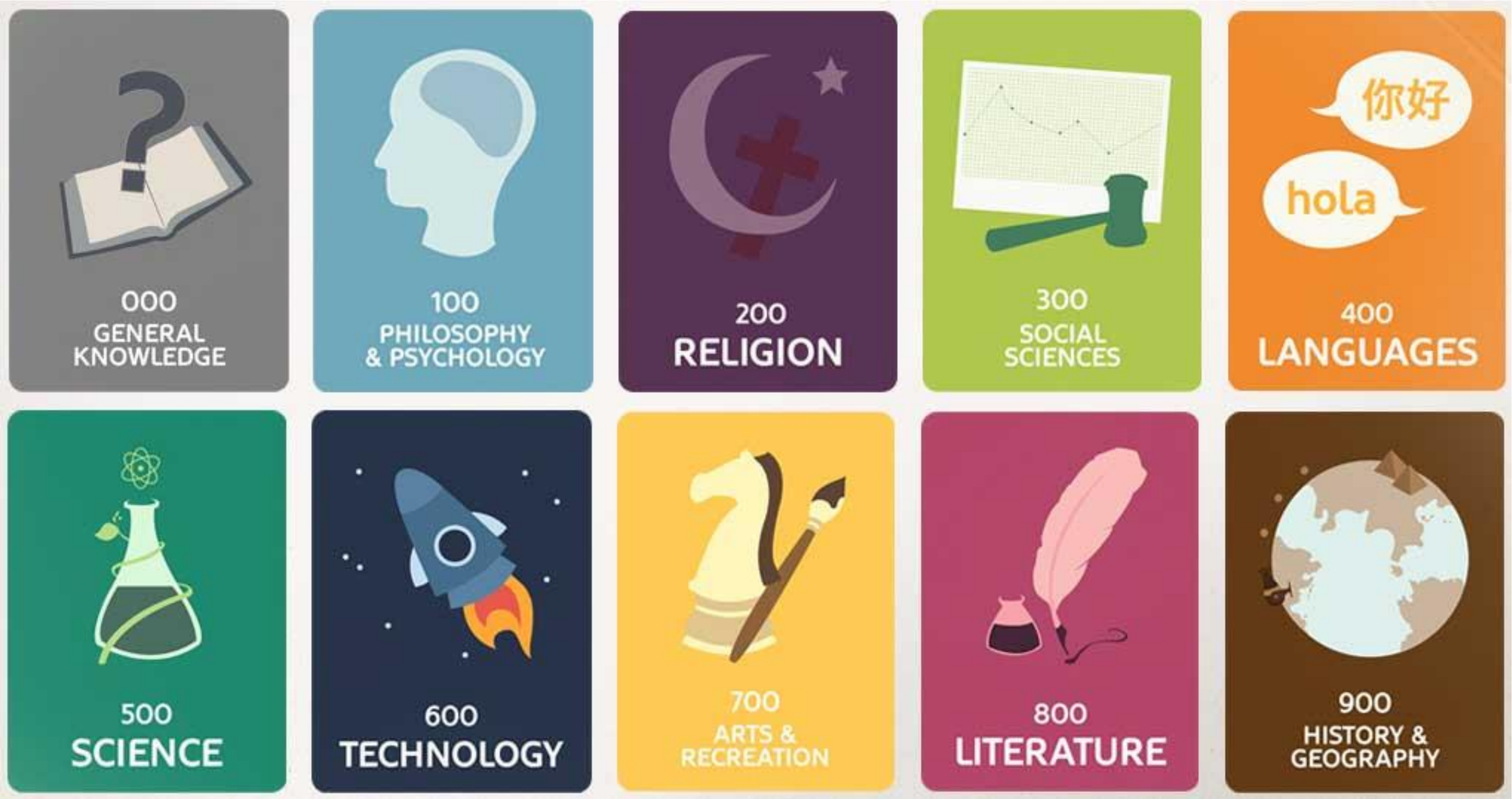
“***Data curation is the organization and integration of data collected from various sources***. It involves annotation, publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation.”
– *Wikipedia*

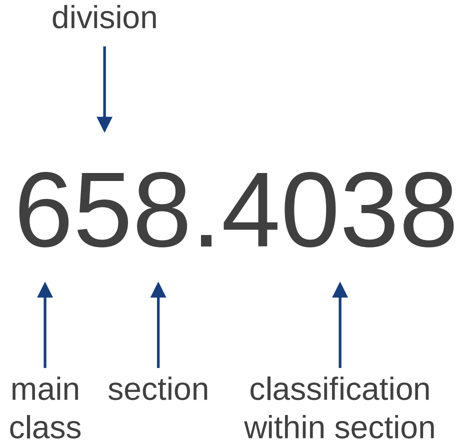
“***Select, organize***, and look after the items in (a collection or exhibition).”
– *oxford dictionary*

Curation Examples

Behaviour in transactions	More complex
The subject/object of an action, e.g., a customer buys a product; a vendor sells a part; a partner delivers a crate of materials to a location; etc.	Simple data consisting of few attributes are usually easily managed, but those data consisting of many and varied dimensions are candidates to be considered
Has a lifecycle	More value
Usually viewed as a CRUD cycle, e.g., for product, • Create – a product gets manufactured • Read – its information is retrieved from an inventory catalogue • Update – changes get made to the product, e.g., packaging, raw materials • Delete – removed due to obsolescence	The more valuable a data element, the more important it is to be considered
Large set of elements	Reusable
The more number of instances of such data, the more important this data is to the organisation, e.g., if there are only two customers, it is unlikely that information on them can be in error, but if there are two thousand customers, then information on them cannot rely on memory alone but must be considered	The more frequent the reuse, the more important it is that the data is considered
Less volatility	Other enterprise-designated critical data
Data that changes infrequently need to be considered, e.g., a mortgage agreement, whereas the mortgage payments are regular and are of interest only for their respective periods of concern	Any data used for compliance or security purposes must be considered. Any other enterprise-designated critical data must be included

Dewey Decimal System




division

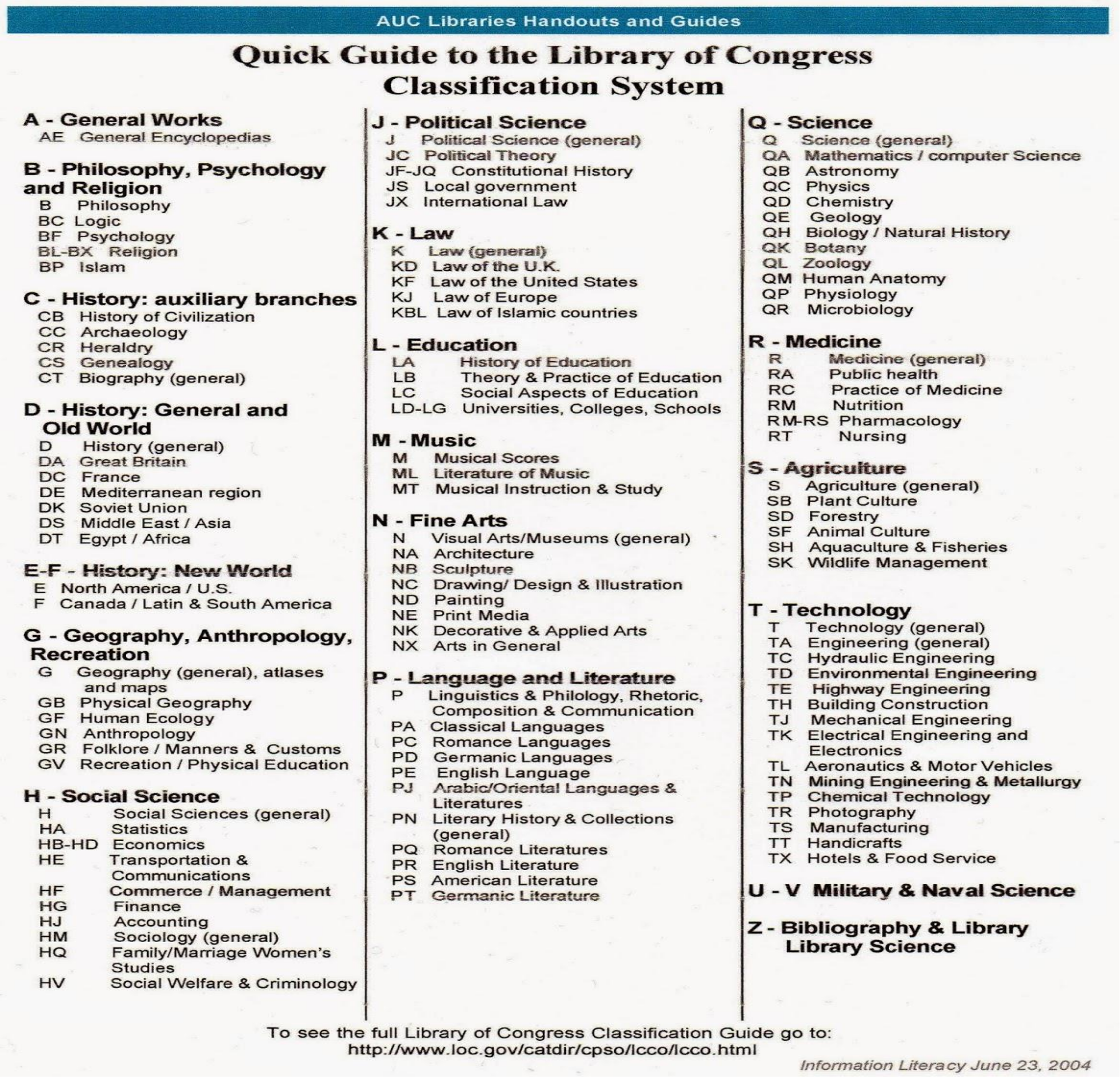
658.4038

main class

section

classification within section

Library of Congress



AUC Libraries Handouts and Guides

Quick Guide to the Library of Congress Classification System

A - General Works

AE	General Encyclopedias

B - Philosophy, Psychology and Religion

B	Philosophy
BC	Logic
BF	Psychology
BL-BX	Religion
BP	Islam

C - History: auxiliary branches

CB	History of Civilization
CC	Archaeology
CR	Heraldry
CS	Genealogy
CT	Biography (general)

D - History: General and Old World

D	History (general)
DA	Great Britain
DC	France
DE	Mediterranean region
DK	Soviet Union
DS	Middle East / Asia
DT	Egypt / Africa

E-F - History: New World

E	North America / U.S.
F	Canada / Latin & South America

G - Geography, Anthropology, Recreation

G	Geography (general), atlases and maps
GB	Physical Geography
GF	Human Ecology
GN	Anthropology
GR	Folklore / Manners & Customs
GV	Recreation / Physical Education

H - Social Science

H	Social Sciences (general)
HA	Statistics
HB-HD	Economics
HE	Transportation & Communications
HF	Commerce / Management
HG	Finance
HJ	Accounting
HM	Sociology (general)
HQ	Family/Marriage Women's Studies
HV	Social Welfare & Criminology

J - Political Science

J	Political Science (general)
JC	Political Theory
JF-JQ	Constitutional History
JS	Local government
JX	International Law

K - Law

K	Law (general)
KD	Law of the U.K.
KF	Law of the United States
KJ	Law of Europe
KBL	Law of Islamic countries

L - Education

LA	History of Education
LB	Theory & Practice of Education
LC	Social Aspects of Education
LD-LG	Universities, Colleges, Schools

M - Music

M	Musical Scores
ML	Literature of Music
MT	Musical Instruction & Study

N - Fine Arts

N	Visual Arts/Museums (general)
NA	Architecture
NB	Sculpture
NC	Drawing/ Design & Illustration
ND	Painting
NE	Print Media
NK	Decorative & Applied Arts
NX	Arts in General

P - Language and Literature

P	Linguistics & Philology, Rhetoric, Composition & Communication
PA	Classical Languages
PC	Romance Languages
PD	Germanic Languages
PE	English Language
PJ	Arabic/Oriental Languages & Literatures
PN	Literary History & Collections (general)
PQ	Romance Literatures
PR	English Literature
PS	American Literature
PT	Germanic Literature

Q - Science

Q	Science (general)
QA	Mathematics / computer Science
QB	Astronomy
QC	Physics
QD	Chemistry
QE	Geology
QH	Biology / Natural History
QK	Botany
QL	Zoology
QM	Human Anatomy
QP	Physiology
QR	Microbiology

R - Medicine

R	Medicine (general)
RA	Public health
RC	Practice of Medicine
RM	Nutrition
RM-RS	Pharmacology
RT	Nursing

S - Agriculture

S	Agriculture (general)
SB	Plant Culture
SD	Forestry
SF	Animal Culture
SH	Aquaculture & Fisheries
SK	Wildlife Management

T - Technology

T	Technology (general)
TA	Engineering (general)
TC	Hydraulic Engineering
TD	Environmental Engineering
TE	Highway Engineering
TH	Building Construction
TJ	Mechanical Engineering
TK	Electrical Engineering and Electronics
TL	Aeronautics & Motor Vehicles
TN	Mining Engineering & Metallurgy
TP	Chemical Technology
TR	Photography
TS	Manufacturing
TT	Handicrafts
TX	Hotels & Food Service

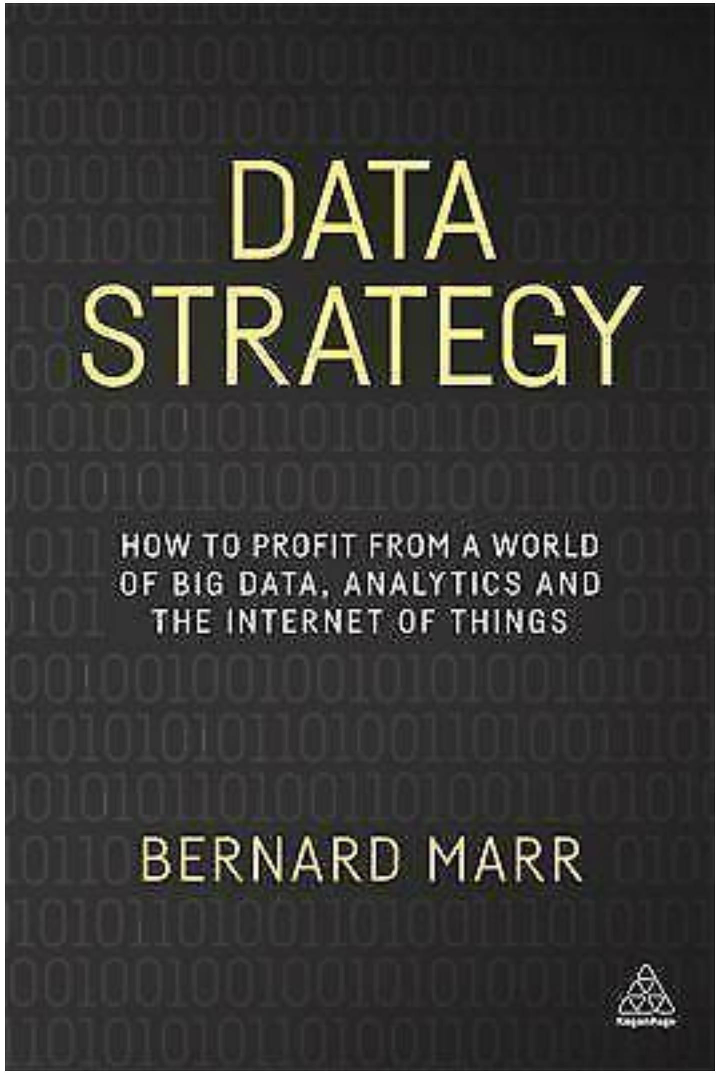
U - V Military & Naval Science

Z - Bibliography & Library Library Science

To see the full Library of Congress Classification Guide go to:
http://www.loc.gov/catdir/cpso/lcco/lcco.html

Information Literacy June 23, 2004

HF54.5


DATA STRATEGY
HOW TO PROFIT FROM A WORLD OF BIG DATA, ANALYTICS AND THE INTERNET OF THINGS
BERNARD MARR
KoganPage

Describing Data

Describe the data itself (e.g., databases, data elements, data models, etc.), the concepts the data represents (e.g., business processes, application systems, software code, technology infrastructure, etc.) and the connects (relationship) between the data and concepts.

Business Description	Technical	Operational
Data models, definitions and descriptions of data sets, tables and columns	Physical database table and column names, and properties	Logs of job execution for batch programs
Business rules, data quality rules and transformation rules	Data access rights, groups and roles	Results of audit, balance, control measurements and error logs
Data provenance	Data CRUD (Create, Replace, Update and Delete) rules	Reports and query access patterns, frequency and execution time
Data standards and constraints	ETL (Extract, Transform, Load) job details	Patches and version maintenance plan and execution; current patching level
Security/privacy level of data	Data lineage documentation, including upstream and downstream change impact information	Backup, retention, date created and disaster recovery provisions
	Content update job cycle job schedules and dependencies	

Description Example

ID	Business Data Object	Field Name	Description	Alternate Names	Associated Business Data Object	Data Field	Unique Values?	Data Type	Length	Valid Values	Default Value
DD001	Order	Shipping Address	The entire shipping address for the order	Ship-to Address	Address	shipping address	N	Alphanumeric	50	May only contain letters, digits and periods. May not contain a "PO" or "P.O." as a word. Case insensitive	Customer's preferred Shipping Address if returning customer, otherwise null
DD002	Order	Billing Address Same As Shipping	An indicator of whether the shipping address and the billing address are the same		N/A	billing address same	N	Boolean	N/A	True/False	Customer's preferred setting if returning customer, otherwise TRUE
DD003	Order	Billing Address	The entire billing address for the order		Address	billing address	N	Alphanumeric	50	May only contain letters, digits and periods	Customer's preferred Billing Address if returning customer, otherwise null
DD004	Order	Coupon Code	Payment can be made in full or partial with use o valid promotional coupon or codes	Valid system coupon	N/A	payment coupon	N	Alphanumeric	15	Any	null
DD005	Order	Payment Info Subtotal	Subtotal of price of items in cart	Cart subtotal	N/A	payment subtotal	N	Currency	10	0.00…999,999.99	null
DD006	Order	Payment Info Sales Tax	Sales tax added to the order sutotal depending upon customer's location	Sales Tax $	N/A	payment tax	N	Currency	10	0.00…999,999.99	null

Description Example

ID	Business Data Object	Field Name	Calculation	Reqd?	Business Rules	Customer Role	Sales Rep Role	Track Changes?	Owner	Status	Notes
DD001	Order	Shipping Address	N/A	Y	N/A	View, Edit	View, Edit	Yes	Purchase Team	Reviewed	
DD002	Order	Billing Address Same As Shipping	N/A	N	N/A	View, Edit	View, Edit	Yes	Purchase Team	Reviewed	
DD003	Order	Billing Address	N/A	Y	N/A	View, Edit	View, Edit	Yes	Purchase Team	Reviewed	
DD004	Order	Coupon Code	N/A	N	Must be a legitimate coupon code that is still valid (not expired) and not redeemed	View, Edit	View, Edit	Yes	Business SME	Draft	
DD005	Order	Payment Info Subtotal	If a "Dollar Off" coupon code is entered: Sum of the price of all items in the cart minus the coupon amount. If result is < 0, then 0. If a "Percent Off" coupon code is entered: (Sum of the price of all items in the cart) * (100-Percent Off)/100, rounded to the nearest cent (round 0.5 up). Otherwise: Sum of the price of all itmes in the cart	Y	N/A	View	View, Edit	Yes	Finance	Draft	
DD006	Order	Payment Info Sales Tax	Tax calculated using Payment Info Subtotal and Shipping Address; see "tax calculations"	Y	See "tax calculations"	View	View	Yes	Finance	Draft	

Quality Dimensions of Data

- A data quality *dimension* is a measurable feature or characteristic of data
- The term *dimension* is used to make the connection to dimensions used in measuring physical objects, e.g., length, width and height
- Data quality *dimensions* serve as requirements which are initially baselined and then constantly monitored against targets
- In order to maintain the quality of data, an organisation needs to identify *dimensions* which are important to business; these *dimensions* need to be measurable and actionable

Quality Dimensions of Data

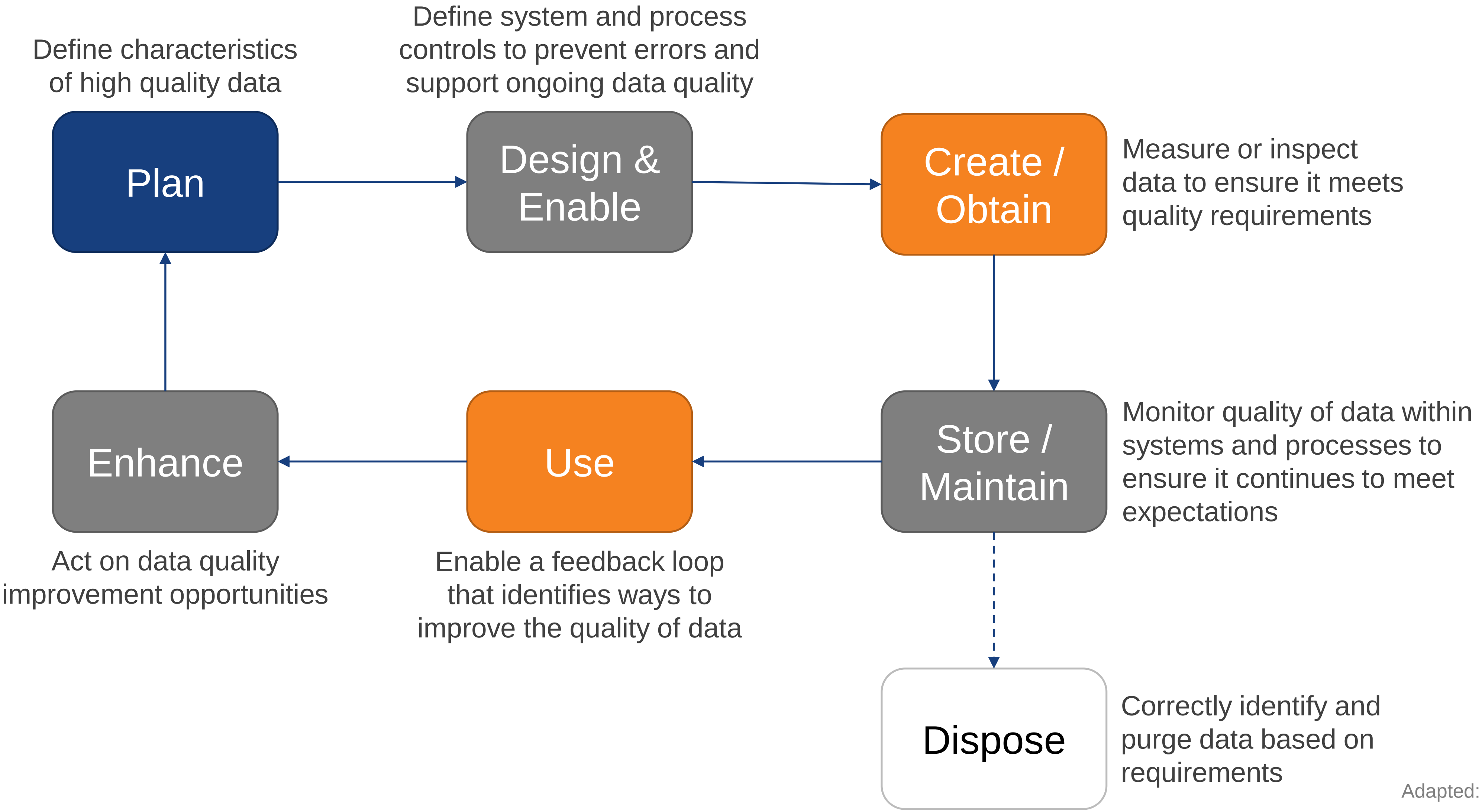
Whilst there is not a single agreed-to set of data quality dimensions, they generally focus on whether there is enough data (completeness); whether it is correct (accuracy, validity); how well it fits together (consistency, integrity, uniqueness); and whether it is up to date (timeliness), accessible, usable and secure.

Dimension	Concepts	Dimension	Concepts
Completeness	The proportion of data stored against the potential for 100%	**Integrity**	The accuracy and consistency of data over its lifecycle
Accuracy	The degree to which data correctly describes the “real world” object or event being described	**Uniqueness**	No entity instance (thing) is recorded more than once based on how the thing is identified
Validity	Data is valid if it conforms to the syntax (format, type and range) of its definition	**Timeliness**	The degree to which data represents reality from the required point in time
Consistency	The absence of difference when comparing two or more representations of a thing against a definition		

Quality Measures

Dimension and Business Rule	Measure	Metrics	Status Indicator
Completeness Business Rule 1: Population of field is mandatory	Count the number of records where data is populated, compare to the total number of records	Divide the obtained number of records where data is populated by the total number of records in the table or database and multiply it by 100 to get to percentage complete	Unacceptable: Below 80% populated; Above 20% not populated
Example 1: Postal Code must be populated in the address table	Count populated: 700,000 Count not populated: 300,000 Total count: 1,000,000	Positive measure: 700,000/1,000,000*100 = 70% populated Negative measure: 300,000/1,000,000 *100 = 30% not populated	Example result: Unacceptable
Uniqueness Business Rule 2: There should be only one record per entity instance in a table	Count the number of duplicate records identified; report on the percentage of records that represent duplicates	Divide the number of duplicate records by the total number of records in the table or database and multiply it by 100	Unacceptable: Above 0%
Example 2: There should be one and only one current row per postal code on the Postal Codes master list	Count of duplicates: 1,000 Total Count: 1,000,000	10,000/1,000,000*100 = 1.0% of postal codes are present on more than one current row	Example result: Unacceptable
Timeliness Business Rule 3: Records must arrive within a scheduled timeframe	Count the number of records failing to arrive on time from a data service for business transactions to be completed	Divide the number of incomplete transactions by the total number of attempted transactions in a time period and multiply by 100	Unacceptable: Below 99% completed on time; Above 1% not completed on time
Example 3: Equity market record should arrive within 5 minutes of being transacted	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000	Positive: (1,000,000 – 2000) / 1,000,000*100 = 99.8% of transaction records arrived within defined timeframe Negative: 2000/1,000,000*100 = 0.20% of transactions did not arrive within defined timeframe	Example Result: Acceptable
Validity Business Rule 4: If field X = value 1, then field Y must = value 1-prime	Count the number of records where the rule is met	Divide the number of records that meet the condition by the total number of records	Unacceptable: Below 100% adherence to the rule

Data Quality through the Lifecycle


Define characteristics of high quality data
Define system and process controls to prevent errors and support ongoing data quality
Plan
Design & Enable
Create / Obtain
Measure or inspect data to ensure it meets quality requirements
Enhance
Use
Store / Maintain
Monitor quality of data within systems and processes to ensure it continues to meet expectations
Act on data quality improvement opportunities
Enable a feedback loop that identifies ways to improve the quality of data
Dispose
Correctly identify and purge data based on requirements

Adapted: DMBOK2

THANK YOU

nicholas_tan@nus.edu.sg