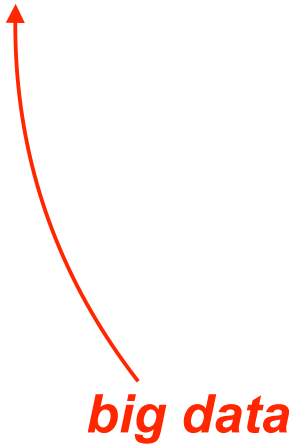
Introduction to Data Management



Week 1

STQD 6324

An Introduction to Data Management

Data

Data Management -> need to properly manage the data [source]

oil & gas

- What is Data? Is Data the New Gold?
- What is the highest orders of magnitude for Data known today?

Multiple-byte units V·T·E

Decimal			Binary				
Value	Metric		Value	IEC		Legacy	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte		–
1000^6	EB	exabyte	1024^6	EiB	exbibyte		–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte		–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte		–

Orders of magnitude of data

need to properly cite the source

https://bit.ly/388LO4A

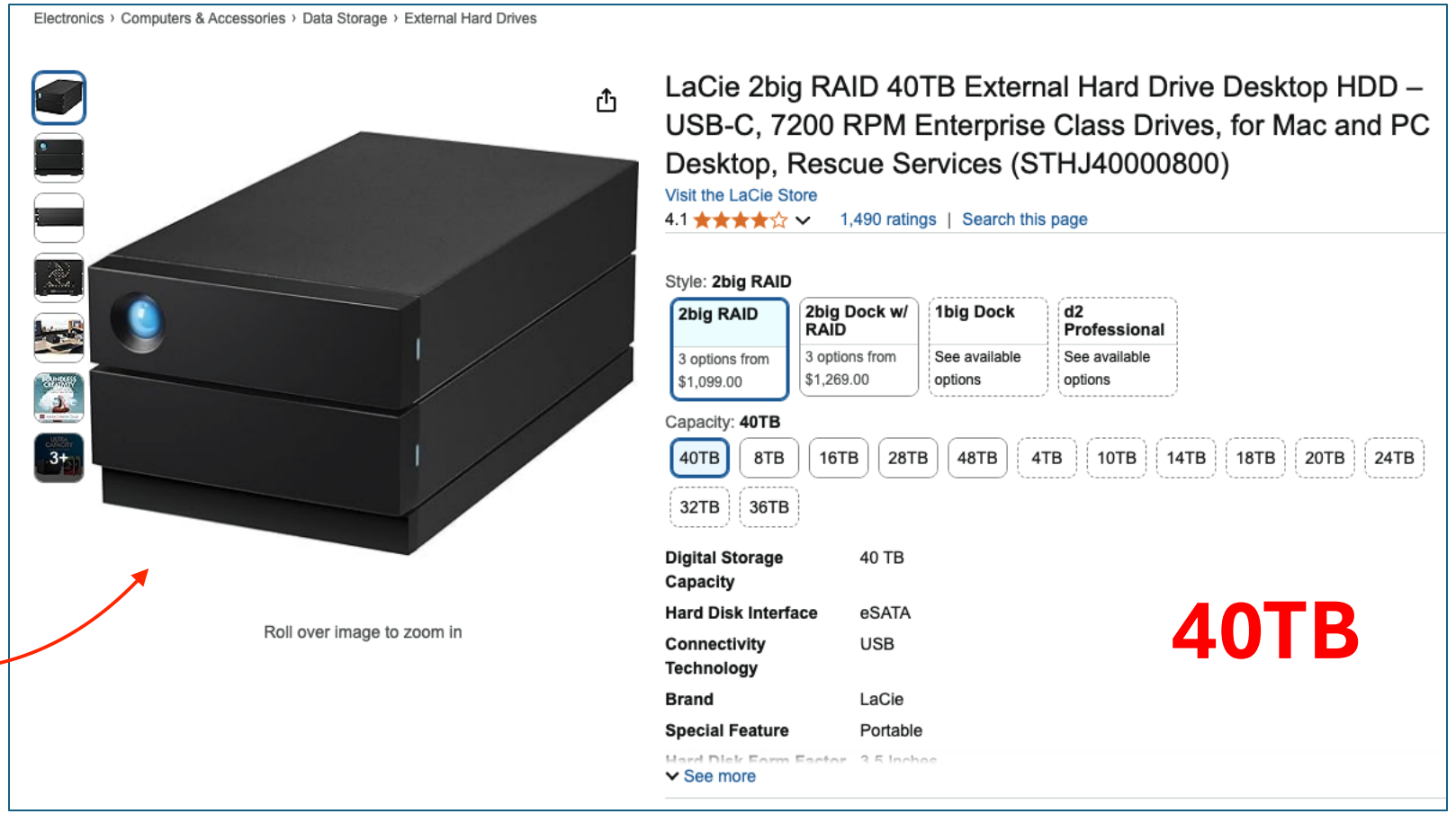
Zettabyte vs. Yottabyte

Feature	Zettabyte (ZB)	Yottabyte (YB)
Storage Size	1 ZB = 10^{21} bytes	1 YB = 10^{24} bytes
Equivalent to	1 billion TB	1 trillion TB
Usage	Cloud storage, Big Data, AI	Future quantum computing, ultra-big data
Real-World Example	Global internet traffic per year	Not yet practically used
Analogue	• A single zettabyte could hold approximately **250 billion DVDs** worth of data • If you were to store one zettabyte of data on standard 4TB hard drives, it would require over **250 billion drives**. • If you were to count to a zettabyte at a rate of one number per second, it would take over **31 billion years to reach one zettabyte**.	

properly manage the data

High-capacity External Hard Disk

Highest capacity on the market

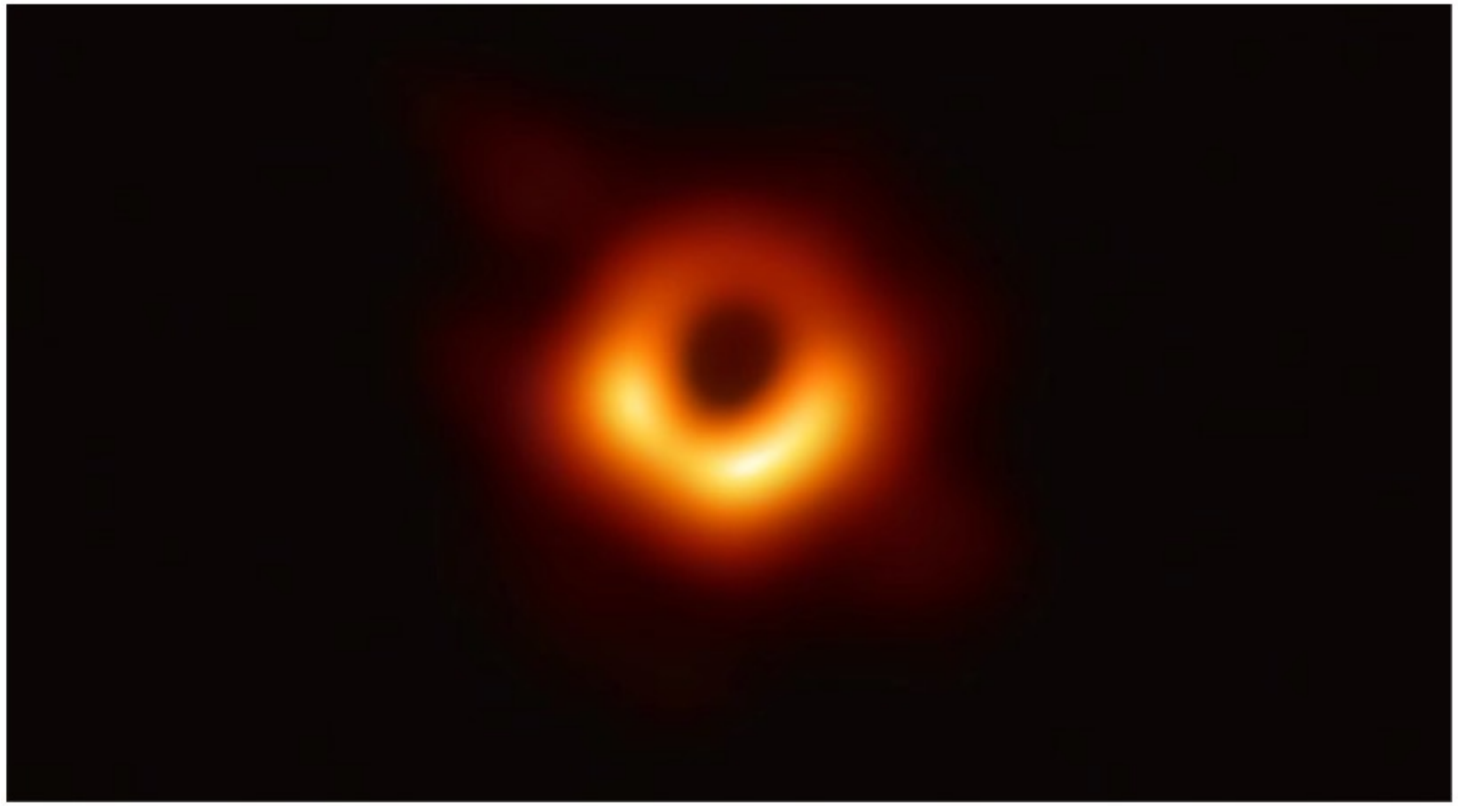


amazon

https://tinyurl.com/yvvaj3h3

2025-03-19

The black hole image data [2019]



https://bit.ly/3DiWBoi

- 5 petabytes **of image size**
- ~ 5,000 years of MP3 audio
- 7 days **[Data Collection]**

• taken by multiple radio telescopes around the world

took over 200 researchers two years to process and analyze the data to create the final black hole image!!

Data Management

- involves the **systematic handling**, **storage**, **processing**, and **sharing of data** in a scientific context to ensure accuracy, accessibility, and reproducibility.
- **FAIR principles (Findability, Accessibility, Interoperability, and Reusability)** to enhance the utility of research data (*Wilkinson et al., 2016*)
- Scientific data are increasingly recognized as critical research assets, requiring proper stewardship, governance, and ethical considerations to maximize their impact. (*OECD, 2021*; *DataCite, 2023*)

Data Management

- Research data are considered all information
  - Collected
  - Observed or created for purposes of analysis and validation of original research results.

- Data can be quantitative or qualitative and comprises also photos, objects or audio files resulting from as different sources as field experiments, model outputs or satellite data.

- The reason data management is important is that the value of research data is sometimes not yet visible nowadays, which can lead to neglecting proper data management.

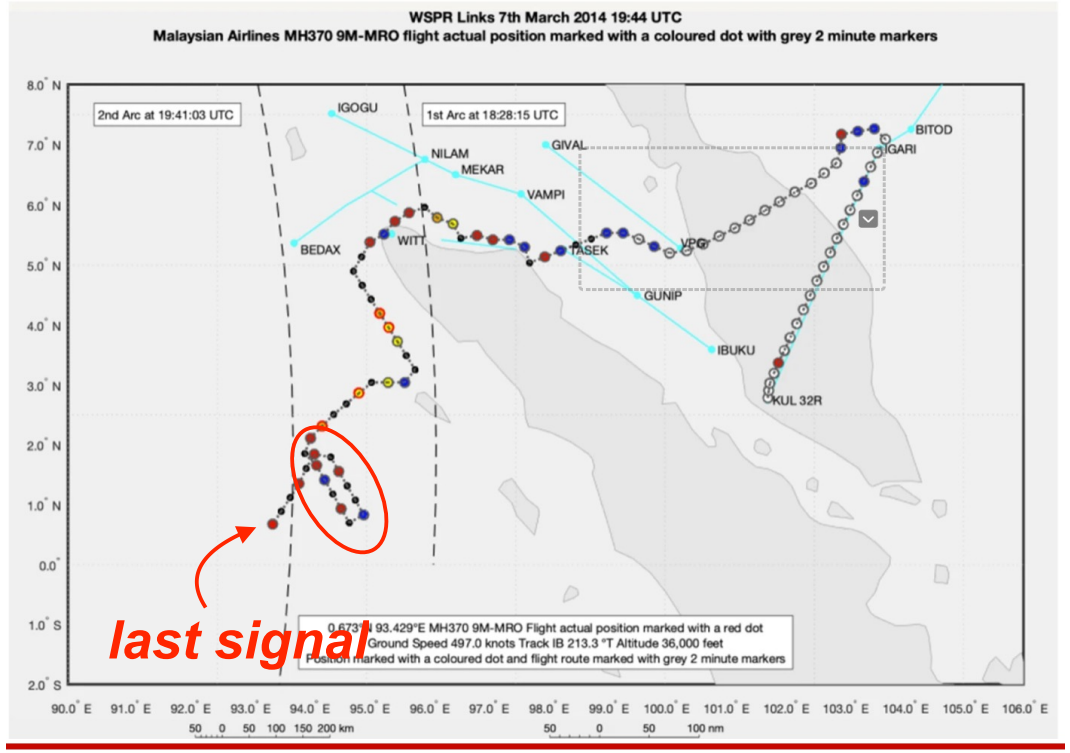
due to technical, knowledge & skill set constraint

MH370

imagine if we didn't properly manage the data in 2014

map the flight path

https://on.ft.com/36WWFhv



last signal

https://bit.ly/3j00OEd

2014 ⟷ **2021**

7 years later

Satellite signal - Doppler effect (an hour)

Radio signal (every two minutes)

gap is closing: an hour to 2 minutes

Doppler effect

- When the observer is **moving towards the source of the wave**, the **frequency or wavelength of the wave appears to increase**.
- Conversely, when the observer is **moving away from the source**, the frequency or wavelength of the wave appears to decrease.
- As the ambulance **approaches an observer**, the frequency of the siren appears to **increase**, and as it **moves away**, the frequency appears to **decrease**.
- Satellite signal – every hour [2014]
- **Radio signal – every two minutes [2021]**

AirlineRatings.com

Airlines Reviews Compare Articles More

MH370 Search Officially ON!

06 March, 2025 2 min read

Airline News

Passenger News Safety

Sharon Petersen

Enter your email Join our newsletter

By joining our newsletter, you agree to our Privacy Policy

Share this story

https://tinyurl.com/mrybkjjs

2025-03-06

Data Management

clinical trial for new drugs
test on patients
- Q: would you like to be
the first patient to test the drugs

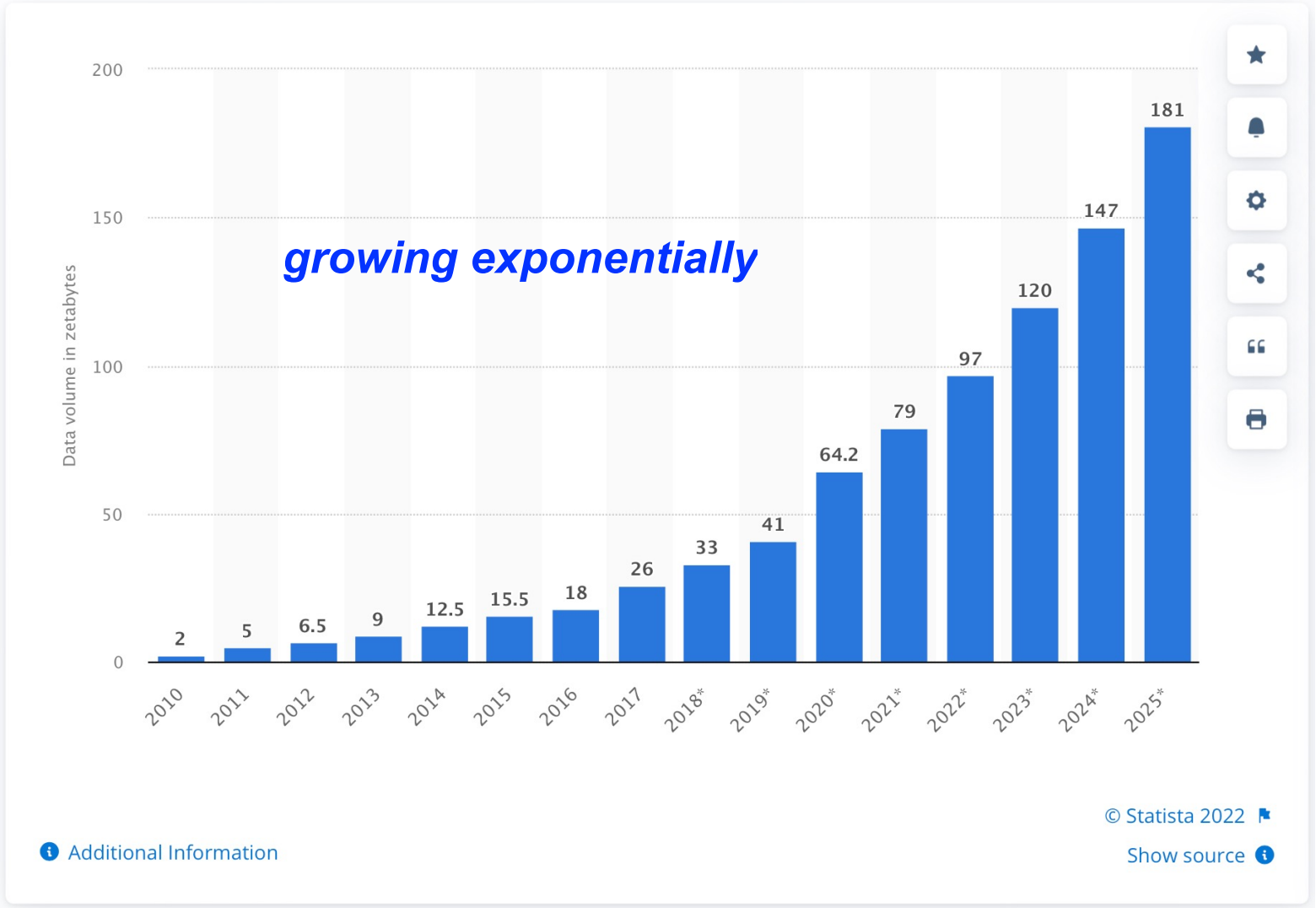
In many sciences experiments or observations cannot be repeated making at least part of the data so valuable that it needs to be stored for a long time. In many cases the value of data can only be realized after many years by new generations (RDA Europe 2014a).

biological data

- One factor making data management even more important is the growing amount of digital data available:

The amount of data collected is growing exponentially nowadays. New environmental observing systems […] will provide access to data collected by aerial, ground-based and underwater sensor networks encompassing tens of thousands of sensors that, when combined, will generate terabytes to petabytes of data annually (Michener and Jones 2012).

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes)



https://bit.ly/3DjNDqT

In 2020:

- **Streaming Netflix for over 34,739 years**
- **Storing all the books in the US Library of Congress over 23 million times**
- **Filling up 59 billion 1-terabyte hard drives**

Data Management

- Data-intensive research is opening up innovative research possibilities (Mantra et al. 2014). If well-managed, these data can be used in order to answer (new) research questions (Corti et al. 2011).

- It includes the collection phase, the processing and analysis of data, the documentation and preservation.

FAIR principle

- Different data management activities are associated with each step, ensuring a reliable and accessible data for the researchers work as well as facilitating sharing and publishing of data.

reproducibility

Data Management

Well-managed data will further facilitate

1. re-use by oneself or others over time,
2. to replicate or validate research results (think of the good scientific practice obligation)
3. processing of so-called wide databases (integration of many small files of varying syntax) and so-called deep databases (handling of BIG data).

merging the data

10.40am

Github -> it is electronic -> everybody can access -> anything you put in public domain
-> open for discussion, open for critics
-> then you have to be really sure what you are doing, make sure they are of quality works
-> make sure you don't leave it to ChatGPT to analyze the results and conclusion for you

Importance and benefits of Data Management

- An example from Mantra Online Course (Mantra et al. 2014) illustrates what role data management can play and how it may support your research:

> *You have completed your postgraduate study with flying colours and published a couple of papers to disseminate your research results. Your papers have been cited widely in the research literature by others who have built upon your findings. However, three years later a researcher has accused you of having falsified the data.*

how to deal this?

thesis -> electronic data; everybody can get hold of it

- Do you think you would be able to prove that you had done the work as described? If so, how?
- What would you need to prove that you have not falsified the data?

How to protect ourself? -> proper data management -> if you properly documented your steps - no one would accuse you of falsifying the result

Importance and benefits of Data Management

- The documentation of data analysis and transformation as well as the storing of data and research results are integral part of data management and could help you to prove your work.

- Without data management, not only a solid basis ensuring replicability for your research results may be missing, but your data can also be subject to data loss more easily.

- This may happen due to technical problems (hardware failure), due to software obsolescence, due to missing information (data cannot be understood in the future) or due to not storing data in an appropriate way (data will never be found again)

Importance and benefits of Data Management

- In Figure 1, the loss of information content of data is related over time to the career of a researcher.
- It illustrates that often much information is tied to specific persons.
- If they leave the project or retire, their knowledge is not available anymore and people do of course also forget details over time.
- Data management may help and facilitate research.

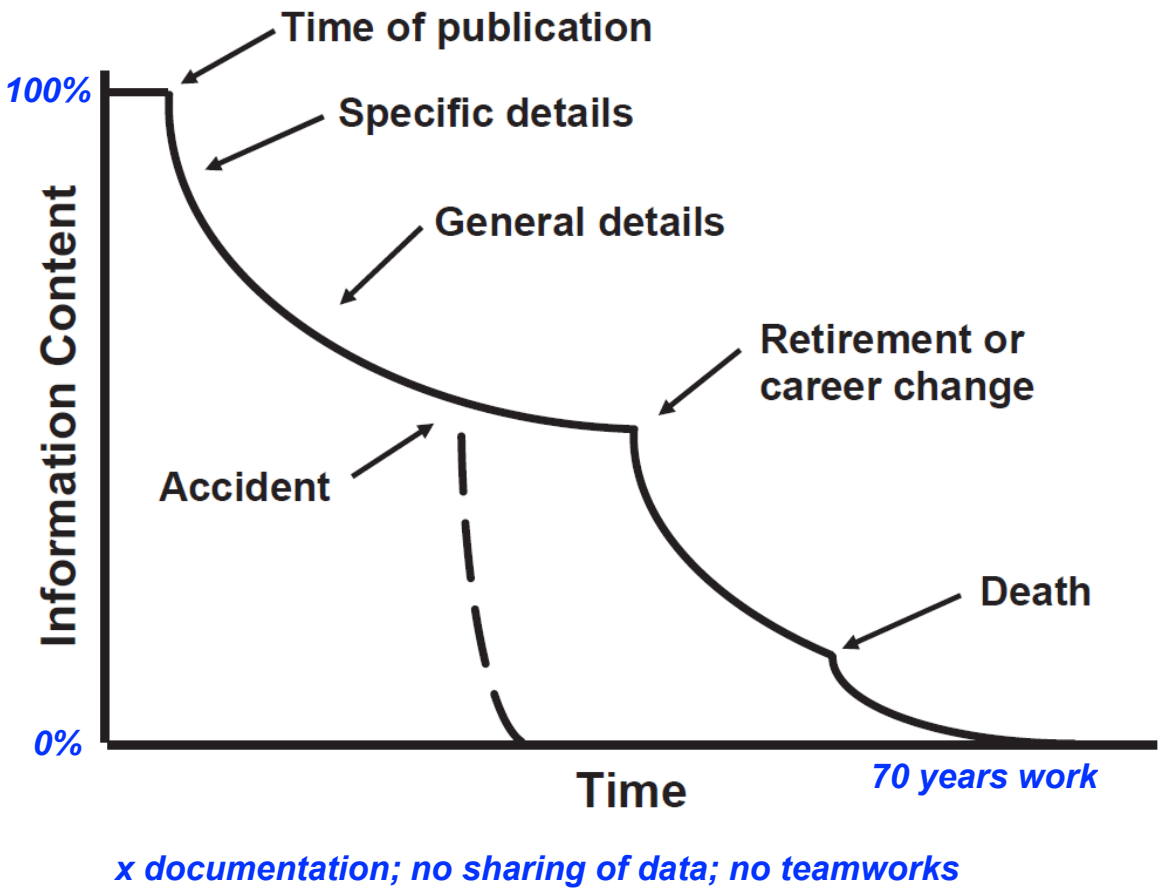


Figure 1: Loss of meta information over time. Michener 2006

Importance and benefits of Data Management

- Data management planning enhances the **security** of data. It safeguards against data loss as storage, backups and archiving are planned.
- **Compliance** to funder or publisher requirements of the collected data is ensured. *share the data*
- **Quality** of research in general is enhanced as data management ensures that research data and records are accurate, consistent, complete, authentic and reliable. It also allows for reproducibility of results.
- Data management planning streamlines data handling and can thus create **efficiency** gains for the whole research project.

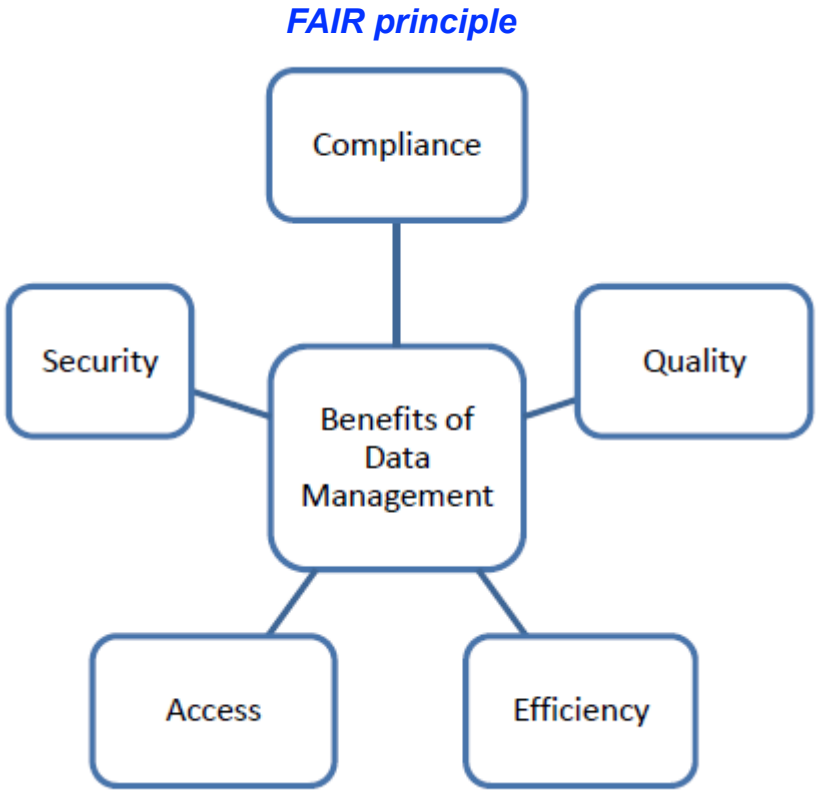


Figure 2: Benefits of Research Data Management Planning. University of Western Australia. (2015)

Importance and benefits of Data Management

- Incorporating data management as a routine part of the research process can save time and resources in the long run.

- In the beginning, some time is needed to prepare a Data Management Plan and to get used to new practices and activities. This is rewarded by extra funding for your data management, increased citations, and less work organising and understanding data later on (DataONE 2012a).

Importance and benefits of Data Management

- Some costs and benefits of data management can be measured quantitatively, in terms of people's time or costs of physical resources like hardware or software (see Figure 3).

- Others have qualitative character or are impossible to measure at all in advance (e.g. possible new scientific findings; Houghton 2011).

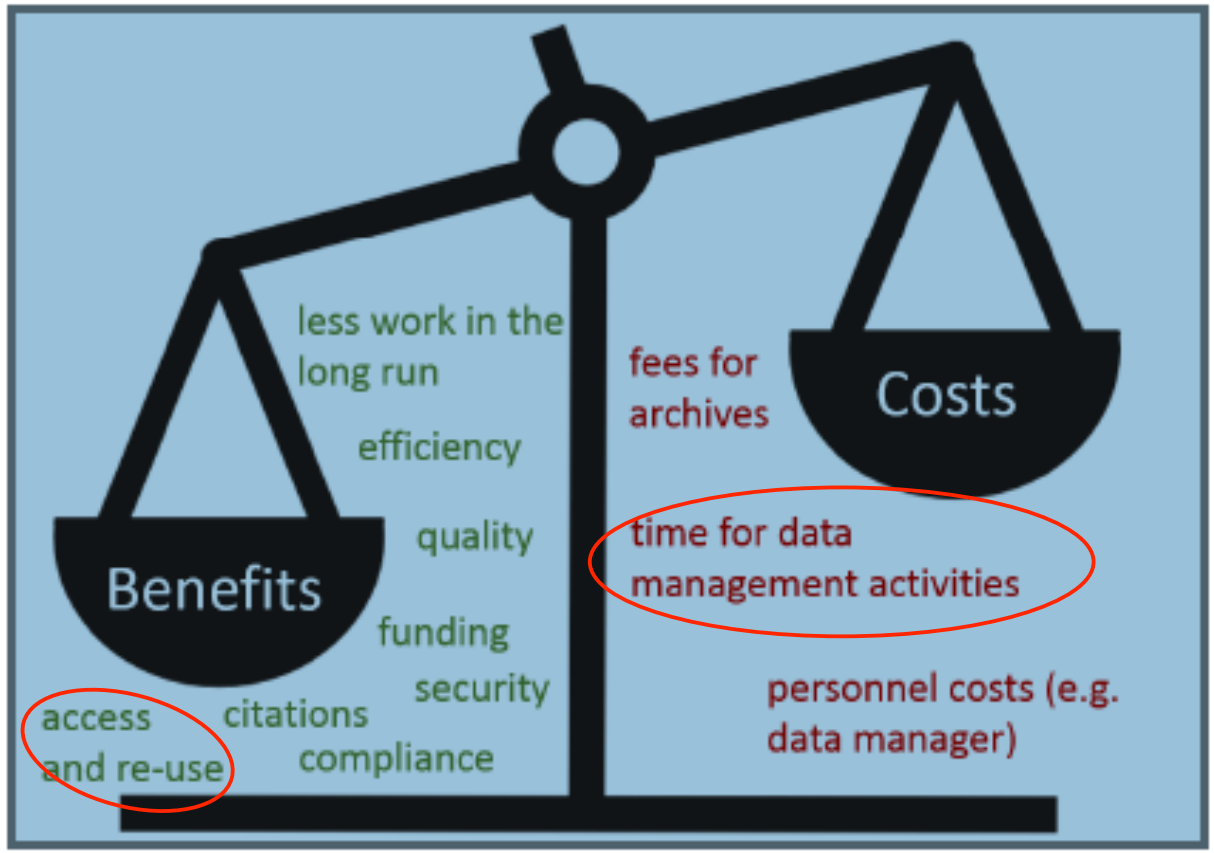


Figure 3: Costs and benefits of data management

Infringement data regulatory compliance

Cases	What happened
• 2018: Facebook was fined £500,000 by the UK Information Commissioner's Office (ICO)	• failing to protect user data in relation to the Cambridge Analytica scandal
• 2019: Google was fined €50 million by the French data protection regulator (CNIL)	• violations of the EU's General Data Protection Regulation (GDPR)
• 2019: Capital One, a US-based bank, was fined $80 million by the Office of the Comptroller of the Currency (OCC) in the US	• suffered a major data breach that affected over 100 million customers

poor data management == cost $$$$$

- cases where involved sensitive data -> like IC number

Advantage of Sharing

- Sharing data facilitates the collaboration within and outside research projects and establishes links to the next generation of researchers because data are discoverable and understandable.
- It also allows to approach research questions which were not thought of when the research started.
- Prevention of unnecessary duplication of data collection.
- A key factor for science is replicability, so researchers can collect data and analyse them in order to produce similar results or assess previous work in the light of new approaches (e.g. voice recording of bat signals and the determination of species) (UK Data Service 2012-2015b).

Data Life Cycle

- The different activities concerning data management can be structured in the so called Data Life Cycle (Figure 4).
- The Data Life Cycle is a conceptual tool which helps to understand 10 different steps that data management follows from data generation to knowledge creation.
- Many steps of the Data Life Cycle are not only performed once, but multiple times or continually over the life cycle.
- The order of the Data Life Cycle is also adapted to the needs of a research project. It can be approached from different perspectives, such as data producer and data re-user.

Figure 4: Data Life Cycle after GFBio

For example, a data re-user does not collect data, and not every data producer integrates data from other researchers.

Data Life Cycle

- Some practices and steps are normally carried out by specialists called (digital) curators working at data repositories.
- Curation is managing digital items in a storage to ensure long-term preservation.
- One step further is to make them discoverable and accessible as soon as it is possible.
- For a short overview over the different steps of the Data Life Cycle, the fact sheets on the GFBio Homepage are recommended (http://www.gfbio.org/data-life-cycle).

Data Life Cycle

Capstone project

- The **Research Life Cycle** (Figure 5) is a model for the steps followed in order to create scientific knowledge.

- The Research Life Cycle, depicted in orange, starts with the research idea and comprises the establishment of cooperation with research partners, the composition of a research proposal, the granting of funding, the Data Life Cycle (except the step "propose", which is already included in the research proposal) and finally the publication of research results.

- The generated scientific knowledge and information serves as starting point for new research problems.

- align with the company KPI
- company wish to accomplish

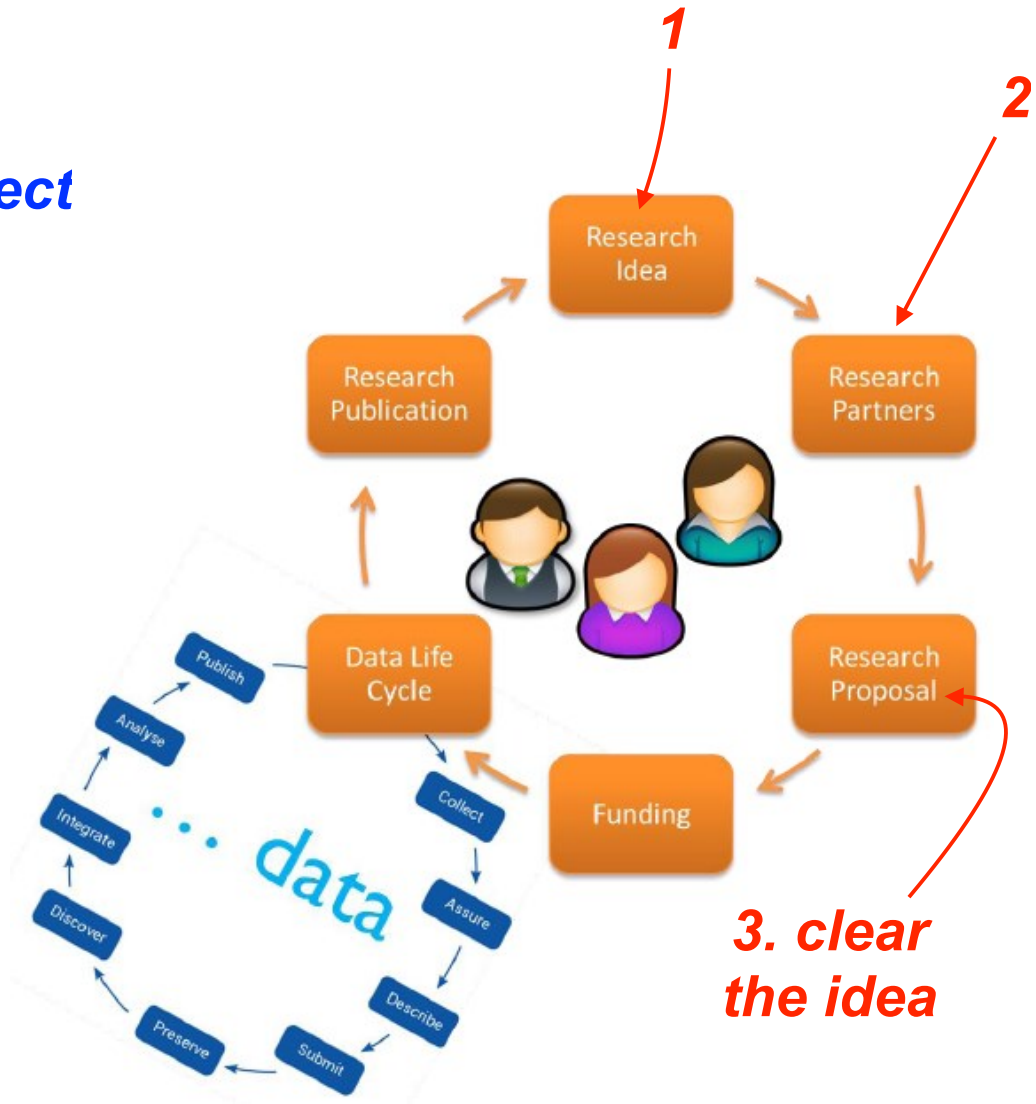


Figure 5: Research Life Cycle after GFBio

1: Propose

discussion;
things that you think is interesting,
may be others does not like it;
don't syok sendiri

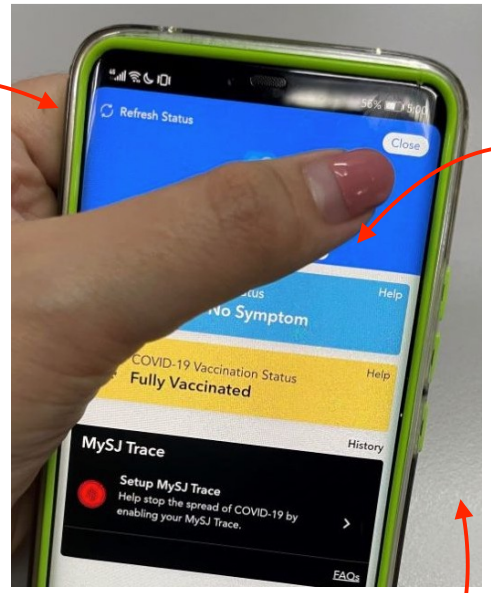
- Data management ideally begins at the ***planning and proposal*** phase of the research project.

Assemble your team members

- This is the best moment to establish a Data Management Plan to provide a framework that supports researchers and their data throughout the course of research and to provide guidelines for everyone to work with (Mantra et al. 2014).
- Table 1 gives a first insight in what can be included in a Data Management Plan.

Poor data management

MySejahtera App – Who Owns what?



Entomo Malaysia

MySJ

Malaysian government

MySejahtera sensitive data -> in Server in Singapore

PDPR

Interesting info:

- 38 million users
- 1 million businesses in the last 5 months **[2022-03-28]**
- "Software as a service"
- Until December 31, 2025
- RM338.6 million

worth

https://tinyurl.com/4d2tv8cn

https://bit.ly/3uysi9b

1:Propose

- Establishing a Data Management Plan at proposal stage in the Data and Research Life Cycle facilitates a structured work with data and saves time later on.
- A Data Management Plan is a living document that is to be maintained and kept up-to-date, e.g. if staff changes.
- It is important to base the plan on available resources and support to ensure that implementation is feasible.

2: Collect

data -> Q: Why do I want to collaborate with you

data == $$$$

- Collection includes various procedures such as ***manual recordings of observations*** in the laboratory or field on hand-written data sheets as well as automated collection by data loggers, satellites or airborne platforms (Michener and Jones 2012).
- When collecting data, it is helpful to think of subsequent steps of the Data Life Cycle: ***what is going to happen with the data?***

break - come back at 11.45am

2: Collect

- Here are some tips for data collection which are particularly important when several people are involved with data collection and entry:
  - Decide what data will be created and how - this should be communicated to the whole research team.
  - Be clear about methods.
  - Use collecting protocols.
  - Develop procedures for consistency and data quality.

Capstone -> if it involved sensitive data; you can mask the results in the thesis

2: Collect : Data Entry

tsv: bernard\t18\tbangi\t50000
csv: bernard,18,bangi,50000

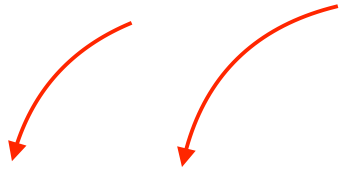
- Also, the file format for working with the data can differ from the formats used for storing and long-term preservation.
- Excel can be a good choice for data entry, but use a syntax that allows information to be stored without loss in csv-files (so that they can be easily accessed with other programs e.g. for analysis).

csv - comma separated value
https://bit.ly/3JTSUYD

- If spreadsheets are used, systematic and accurate work from the beginning on facilitates the process of data exchange to other programmes like R.
- Especially for analysis and data transformation, it is recommended to use scripted environments like R.

csv -> comma separated value
tsv -> tab separated value

2:Collect : Data Entry



- For entering data, it is recommended to use codes like ASCII, UTF-8 or ISO 8859-1 (Latin1). These codes contain characters which can be read by most programmes without any problems (in contrast to codes using e.g. umlauts).

Youtube explanation: https://bit.ly/3iJss8k

- If a file is opened using a wrong encoding, something like this can happen:

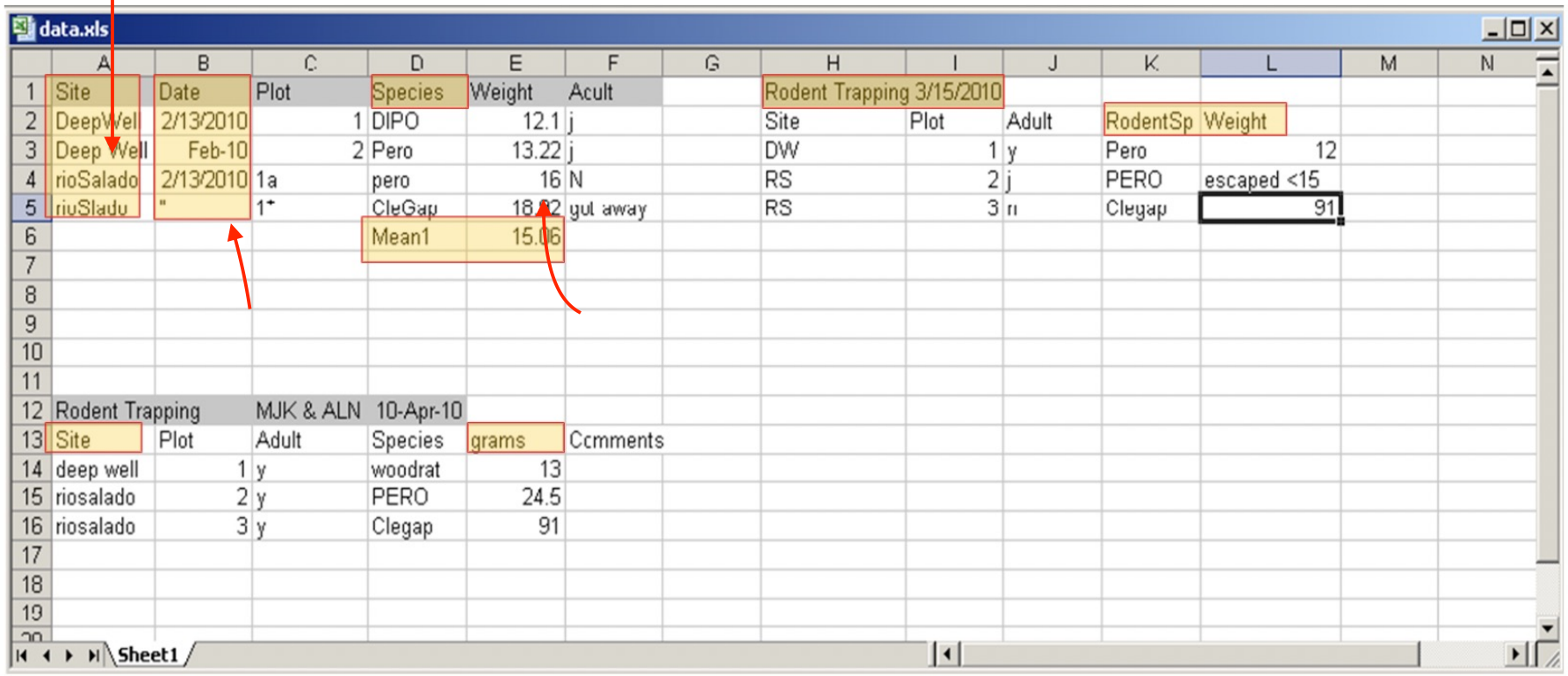
Bärbel Bürgenßen (2015): Encoding Problems in the City of Mörgäl.

ï»¿BÃ¤rbel BÃ¼rgenÃŸen (2015): Encoding Problems in the City of MÃ¶rgÃ¤l.

Poor data entry

put underscore

data cleaning



data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWel	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1*	CleGap	18.82	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

Figure 6: Example for poor data entry. DataONE 2012d

- A table structured in this way is hard to filter and to analyse.
- Data should be ***structured*** as consistent as possible.
- Even if there are any errors, they can be fixed much easier (via scripting) than if the data entry was extremely unstructured.

2: Collect : Spreadsheet Structure

SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

SmallMammalTrapping | Sheet3

Figure 7: Example for good data entry. DataONE 2012d

- The entries are consistent now: only numbers or dates or text was entered.
- Consistent names, codes and formats (date) are used in each column. And data are all in one table, which is much easier for a statistical programme to work with than multiple small tables which each require human intervention.
- This record also underlines the importance of additional information about the data (metadata).
- It is not apparent from the table what measurement unit is used for weight, or what the species abbreviations mean.
- This information can be given on a separate sheet or in the metadata documentation of the dataset

2:Collect : File Naming

-> best: when you saw the file name, you already know what it is

- Naming files is important for organising data on a lab's network drive or personal hard disk and for identifying it later.
- In data repositories the corresponding information is available in the metadata.
- File names should be unique and use ASCII characters and avoid spaces. To ensure file can be read by different operating system and programs.
- The amount and type of information in a file title varies, depending on the type and amount of data and the projects requirements.
- The content of the file should be reflected in the title.

Collect : File Naming

Table 2: Example file naming.

Title	Information content
1. Water samples	???
2 Rhine_water_samples_20140901_V1.0	Rhine (where) water_samples (what) 20140901 (when) V1.0 (version status)
3. Ecoproject_2011_2014_Water_quality_ Rhine_water_samples_Cologne_ 20140901_V1.0	Additional information can be documented in metadata

- The first name has very little information, whereas the third title is already very long.
- Much of that information could be documented in the metadata (project, place, time of collection, time of processing, subject). ***meta -> additional information***
- Including a version number in the file name is a good idea to identify the most recent file, be able to return to older versions and indicate that changes and transformations have been executed on the data.

version number -> just to know which one is new and old data
versioning control

Collect : File Versioning

Action	Version number
I create a new dataset (title does not exist in BExIS).	1.0.0
I upload some data into the dataset.	1.1.0
I make some changes in the metadata (e.g. the address).	1.1.1
I delete some faulty data from the dataset.	1.2.1
The next year, I create a new dataset based upon the dataset I created before.	2.0.0
I upload some data to my newly created dataset.	2.1.0
Etc.	Etc.

Table 3: File versioning in BExIS. BExIS How To: Version numbers in BExIS.

- Including a version number in the file name is a good idea to identify the most recent file, be able to return to older versions and indicate that changes and transformations have been executed on the data.
- In Table 3, an example for a versioning system used by BExIS is given.
- Changes in metadata are indicated by the third digit, smaller changes to the dataset by the second digit and major alterations by the first digit.

3: Assure

- Assure refers to quality control and assurance. Assurance encompasses all those activities which ensure the reliability of data.
- High quality data are a key element for research and impact replicability of results.
- Quality checks should be performed during collection, data entry and analysis and answer the following questions:
  - Are the data complete?
  - Are the data correct?
  - Is the format consistent throughout the data set?
  - If it contains errors, which errors?
  - Are there missing values?

3: Assure

rough idea of numerical data -> scatter plot

- Quality assurance is already applied prior to data collection by defining standards for formats, codes, units and metadata.
- Also the assignment of responsibility for data quality is part of quality assurance (Michener and Jones 2012).
- In the validation process it should be checked whether data are incomplete, unreasonable or inaccurate. This can already be included in the data entry process.
- Statistical and graphical summaries (e.g. max/min, average, range) help to check for impossible values and outliers. After validation, the dataset is cleaned. This means to check outliers, correct and fix errors.

3: Assure : Outliers

outliers -> are the most interesting ones

- Outliers are unexpected values.
- They may not be the result of actual observations, but rather the result of errors in data collection, data recording, or other parts of the Data Life Cycle.
- To identify outliers, statistical tests can be used (Dixon's test, Grubbs test, Tietjen-Moore test).
- Another possibility is the visualization of data (Box plots, scatter plots when there is an expected pattern, such as a daily cycle).
- A third way for detecting outliers is the comparison to related observations.

3: Assure : Outliers

- No outliers should be removed without careful consideration and verification that they are not representing true phenomena.
- Although outliers may be valid observations it is important to identify and examine their validity.
- Outliers may represent data contamination, a violation of the assumptions of the study, or failure of the instrumentation (DataONE n.y. Best Practices: Identify outliers).

4: Describe : Metadata

- Additional information about data is called metadata.
- Metadata describe all aspects of data (e.g. who, why, what, when and where) that would allow one to understand the physical format, content and context of the data, as well as possibly how to acquire, use and cite the data (Michener and Jones 2012).
- Sometimes it can be unclear if a value is considered as metadata or as a record. ***core data (most important)***
- For instance, for one research approach the locational information of an observation or experiment is metadata, whereas in another approach this is “primary” data directly underpinning research results.

5: Submit

Storage purpose *[My PhD: three copies back up; one personal email, one uni email, one external disk]*

Weekly

daily

- Submission is the transfer of data to a curated environment.
- This is usually an archive, a data centre, a repository, or a collection.
- Submission to a curated environment ensures safe long term storage and makes data discoverable for other researchers (in the team or outside).
- During submission phase, researchers can decide on how the data can be accessed.
- Access may vary from immediately available for re-use, available with restrictions or an embargo or not accessible for others.

6: Preserve

external hard disk -> wear and tear; accidentally dropped from table

handle by third party -> AWS, GD, GitHub -> out source the risks to third party

- Digital data are fragile. Hardware fails, software becomes obsolete, files are subject to bit rot which produces bit errors.
- Digital preservation as one part of digital curation encompasses a set of actions which ensure long-term usability by maintaining the accessibility, integrity and longevity of data:
  - Longevity: extend the lifespan of data for current and future user requirements
  - Integrity: ensure the authenticity and reliability of data (no undocumented manipulations)
  - Accessibility: store data in formats which ensure their future use

(Source: DCC 2009, DCC2008d)

6: Preserve : Backup

- Backup of data is not the same as preservation.
- Backups are short-term recovery solutions whereas preservation includes measures taken for long term storage and archiving.
- For backups, there exist ideally at least 3 copies of a file, on at least 2 different media, with at least 1 offsite.
- Managed services like university drives are always a better choice than external hard drives or USB-sticks, but make sure you know the conditions as e.g. normal network drives may keep backup copies of a file for only 6 months and if overwritten with messy data and looked at after 7 months you may run into trouble.
- The IT-Team of your institution might give support and advice.

6: Preserve : Migration *SERVERS*

Upload photos into Google drive; then Q: how to I know it is uploaded complete?

- Migration refers to the transformation of data or other digital material from one format or technology to another (software or hardware).
- The ability to retrieve, display and use the contents is maintained by this action (DCC 2008b).
- Files can be migrated within one software product to a newer version or to other file formats when obsolescence occurs.
- Every migration changes data a little bit. If migrations are carried out multiple times, data can be subject to major changes.

make sure data is completely migrated ->

6: Preserve: Migration

- Because functionality is lost and integrity compromised as a result of migration, care must be taken to strict quality checking procedures, for example, to compare the original bit stream and the migrated bit stream.
- To check data integrity and detect bit errors, ***checksums*** can be used.
- A checksum algorithm calculates a value which can be compared to the value of the original file.
- A widely used algorithm is ***MD5***.

6: Preserve : Refreshment & Emulation

- Refreshment
  - Media refreshment refers to copying data to a new medium after a fixed time (e.g., 5 years) to preserve the bit stream and consider the life span of digital storage media
- Emulation
  - Emulation refers to the development of software that can mimic (obsolete) systems on current and future generation of computers. By emulating applications, operating systems and hardware architecture, older files and programmes can still be used and software can be kept alive
  - process of replicating the behavior of one system within another system.

7: Discover *new discovery -> so that others can continue your "legacy"*

- Discovering data means to search and find data collected by other researchers.
- This data can be used for different purposes like long-time analysis, modelling or comparative studies.
- Data and metadata are made accessible through submission on any kind of shared environment.
- The pre-requisite for discovering data is that the authors are willing to share their data with the research community.

sharing data -> valuable data [put embargo; can only be released after six month]

8: Integrate

- Integration is the merging of multiple datasets from different sources, like your recently collected data with former data from other owners, resulting in a new, bigger dataset.
- You might want to integrate a dataset that you discovered and that fits to your own data in order to verify your results, as a starting point for an integrative study or just to test a new hypothesis for a follow-up study.
- Large-scale ecological studies require the integration of data from different studies and disciplines (e.g. population studies, hydrology and meteorology; Michener and Jones 2012).

9: Analyse

- Analysis comprises the actions used to derive and understand information from data.
- The types of analyses depend on the discipline and on the research questions to be answered.
- Furthermore, the software and hardware used to analyse data also vary.

9: Analyse : Documentation

- Data analysis often consist of several steps.
- Analyses are run, hypotheses modified, data transformed and complemented, and again analysed.
- This complex process can be hard to reproduce if it is not properly documented.
- The documentation of data cleaning, transformation and analysis is called process metadata.
- Related to process metadata is the concept of data provenance, which explains the origins of the data at hand.

reproducibility

Reproducibility

FYP undergraduate student

ANALISIS TREND HARGA KIJANG EMAS
DI MALAYSIA MENGGUNAKAN
KAEDAH SIRI MASA

3.5 PLATFORM PERKONGSIAN ANALISIS DAN KOD PYTHON SERTA R

Kajian ini menggunakan perisian R dan Python untuk membangunkan model dan melaksanakan analisis. Kod Python dan R digunakan untuk membina model latihan dan ujian, maramalkan harga, serta menilai prestasi model menggunakan metrik seperti MSE, RMSE, MAE, dan MAPE. Kod penuh bersama data yang digunakan dalam kajian ini boleh dicapai melalui pautan GitHub berikut: https://github.com/JazminaNurinNatasya.

make sure your work is of good quality

10: Publish

- The last step of the Data Life Cycle deals with the publication of data, especially with datasets linked with the publication of related academic papers.
- It is based on the fact that datasets are a fundamental part of the research process, as important as discussions and conclusions derived from them (AGU 2012):

> *The scientific community should recognize the professional value of such [data] activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications.*

Quiz

- Ensure that data is consistent, trustworthy and is used correctly across an organization **[Data Management & Data Quality]**
- Responsible for validating that data is accurate, complete and consistent **[Data Quality & Data Governance]**
- Regulates proper access, use and protection of data **[Data Governance & Data Security & Privacy]**
- Oversees and coordinates all the subdomains into one unifying structure **[Data Governance & Data Management]**

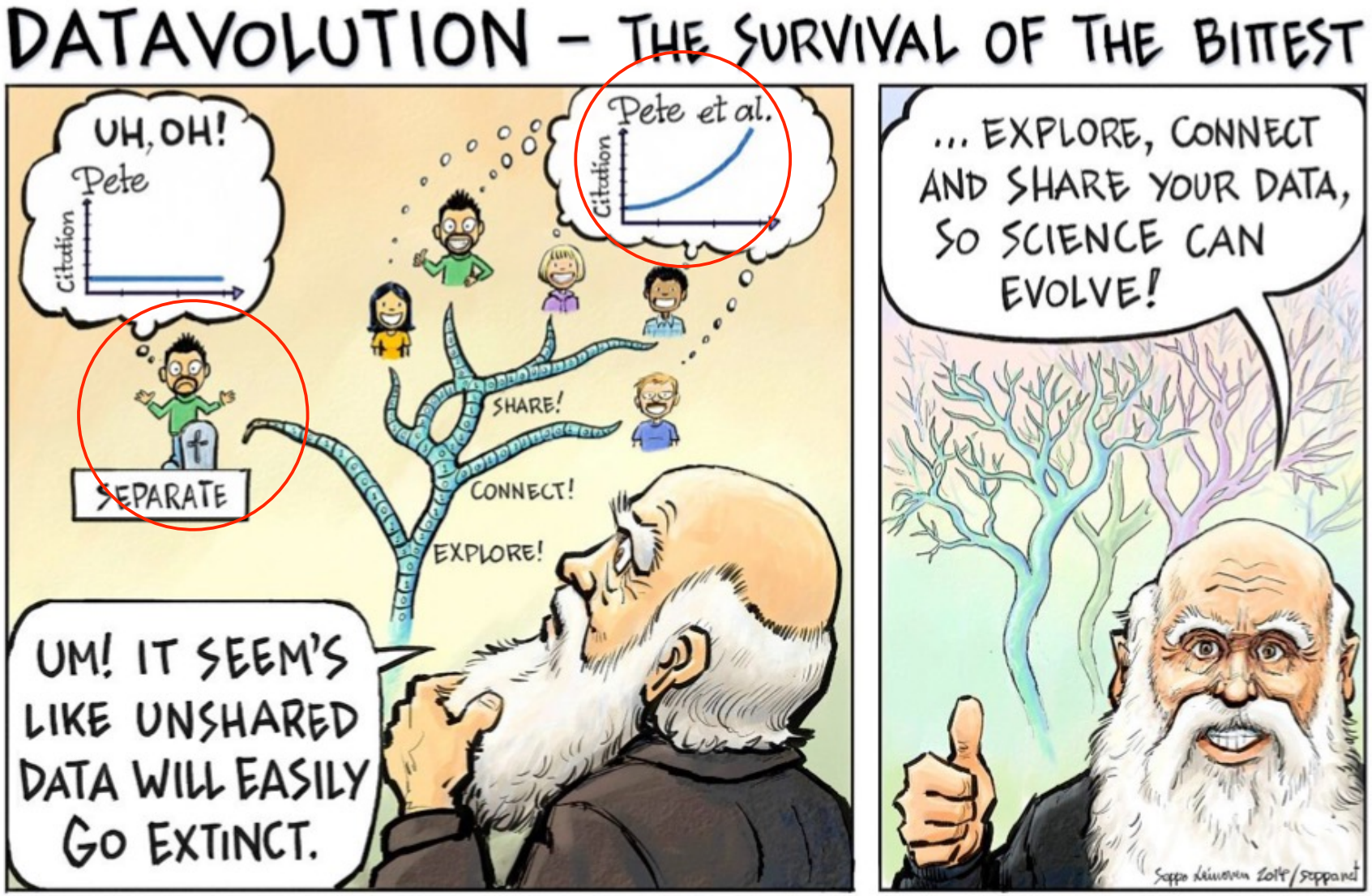


Figure 17: Benefits of sharing data. GFBio-Postcard realised by Seppo

Hadoop installation

- **Requirements**
- 64 bit Windows, MacOSX
- 4GB of RAM for virtual machine (more the better)
- Download and install virtual machine

16 GB RAM; 100 GB of hard disk

2 Hadoop Installation: Oracle VM VirtualBox

1. On PC/Mac

- Download virtual machine: **VirtualBox** (https://www.virtualbox.org)
- Choose the appropriate platform packages
- Require VirtualBox in order to run Hadoop

Windows; Mac

1. ***Install this as well: Visual Studio C++ 2019***

Don't need to buy new laptop or computers yet

3. Horthonworks data platform (hdp) Sandbox

- a *virtual environment* provided by Hortonworks for *learning*, *testing*, and *experimenting* with Hadoop and associated technologies
- allows users to *explore Hadoop ecosystem features* without the need for a full-scale deployment
- HDP Sandbox comes preconfigured with a range of *Hadoop components* and related technologies commonly used in *big data processing and analytics*
- *Hadoop components (you will come across with lots of animals!!!):*
  - *HDFS (Hadoop Distributed File System)*,
  - *YARN (Yet Another Resource Negotiator)*,
  - *MapReduce*,
  - *Hive*,
  - *Pig*,
  - *Spark*,
  - *HBase*

Download hdp sandbox (v. 2.6.5) here: https://bit.ly/3Kd8i53
*** later version no longer an open source**

wait until I give access

start class at 2.30pm