ADVALUE

Advancing Sustainable Agricultural Value Chains through Strengthening Transdisciplinary Skills and Cooperation in East African Doctoral Education

Management and preservation of research data

Hanna Östholm, Swedish University of Agricultural Sciences
Rose Kigathi, Pwani University

11th of August 2025

DATA
MANAGEMENT
SUPPORT

Management and preservation of research data

1. What is research data?
2. Research data management
   a) *Preserve data*
   b) *Publish data*
3. Research data management: Best practice
   a) *Open data and the FAIR principles*
   b) *Top ten tips*
   c) *Document data management*
   d) *Plan data management*

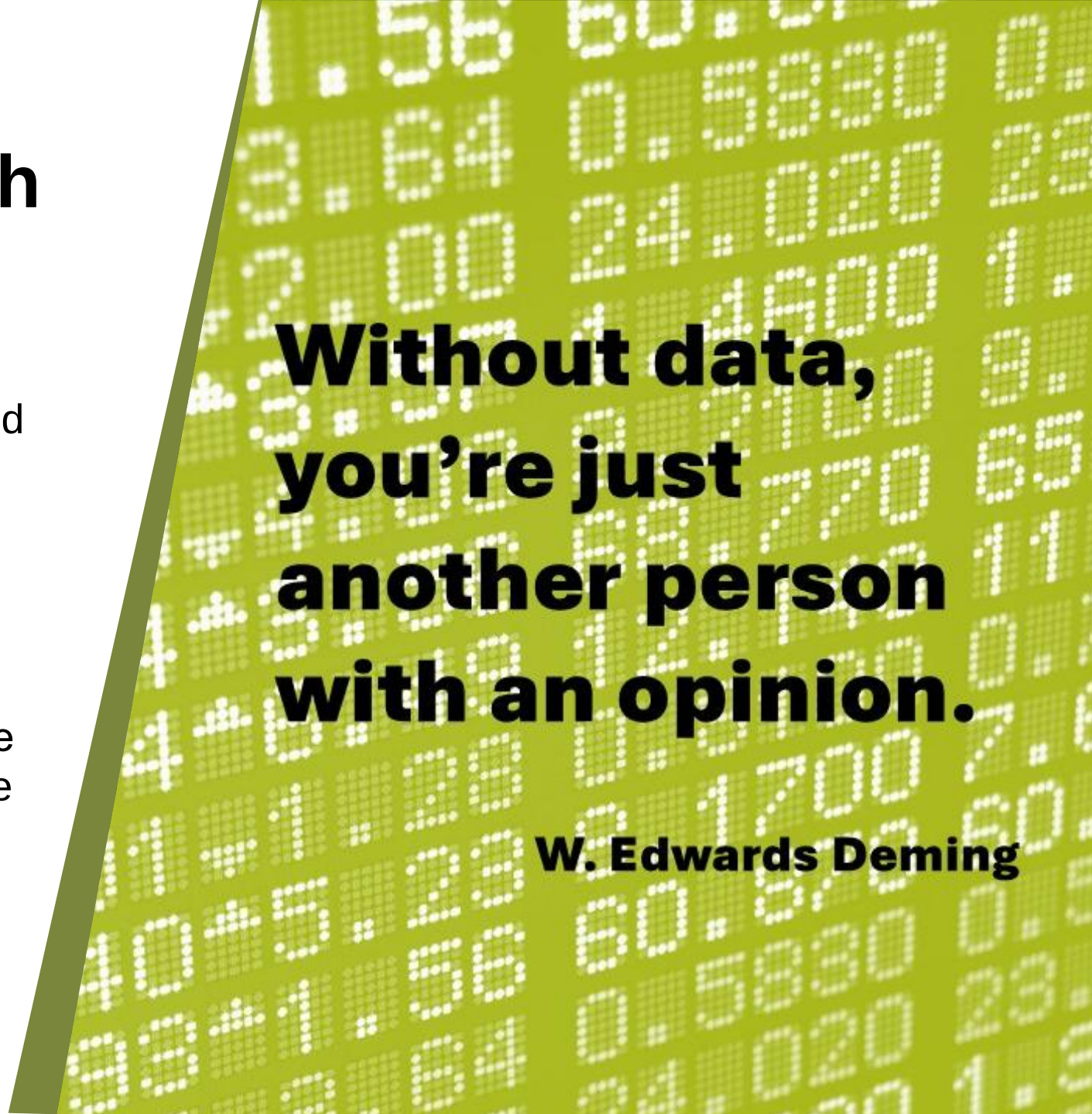
DATA
MANAGEMENT
SUPPORT

What is research data?

DATA
MANAGEMENT
SUPPORT

Definitions of research data

- “Research data refers to information collected to be examined and considered as a basis for reasoning, discussion, or calculation” (from Horizon 2020, p. 4).
- “Research data are the evidence that underpins the answer to the research question, and can be used to validate findings /…/ their identification and value lies in whether and how researchers use them as evidence for claims” (from OpenAire How do I know if my research data is protected?)



dms@slu.se

DATA
MANAGEMENT
SUPPORT

New or reused research data

- Data can be primary – original data collected for a particular purpose/research question, or
- Secondary – data collected for one purpose become secondary when re-used for another purpose.

Illustration: Gerd Altmann, Pixabay (CC BY)

dms@slu.se

Types of research data

- observational
- experimental
- simulated
- derived/compiled
- reference/canonical

Illustration: Gerd Altmann, Pixabay (CC BY)

Data categories

- Qualitative (field notes, interview transcripts, surveys)
- Quantitative (numerical, statistics)

- Numerical (measurements, spreadsheets)
- Descriptive (text documents)
- Spatial (coordinates, images)
- Auditory (audio recordings)
- Visual (image)
- Mixed media (video)

Illustration: Matthew Feeney, Unsplash (CC BY

Stages of research data

- Raw data (from the instrument for measurement or the recording of an interview)
- Analysed or processed data (the version that you will work with, transport into other formats, transcript, quality check, compile or compute for statistics etc.)

Illustration: Gerd Altmann, Pixabay (CC BY)

DATA
MANAGEMENT
SUPPORT

Research data management

Data lifecycle

- Plan data management
- Collect, organise and store data
- Process and analyse data
- Archive and preserve data
- Share and publish data
- Discover, reuse and cite data

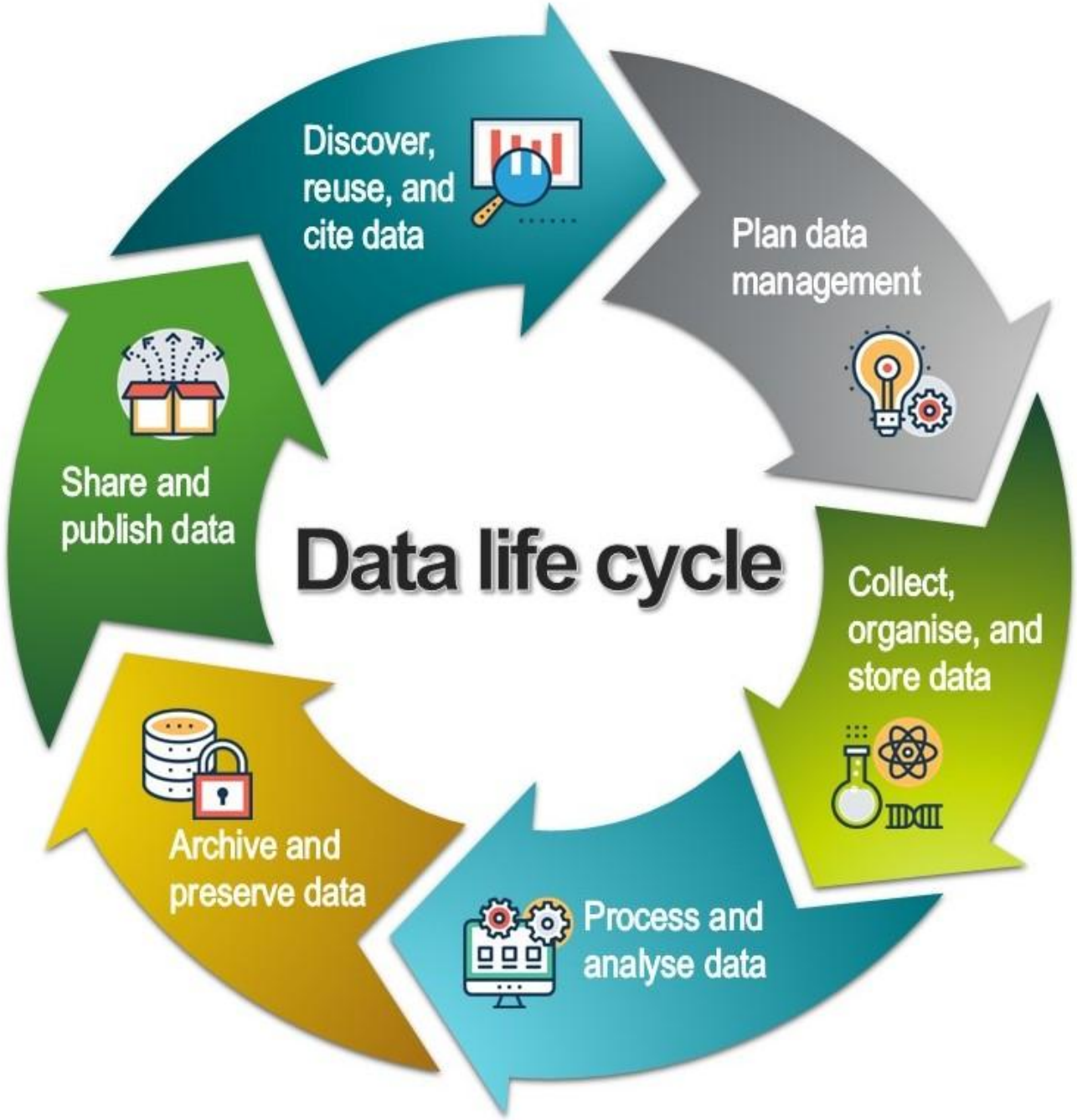


Illustration: Data life cycle by SLU (CC BY)

DATA
MANAGEMENT
SUPPORT

Preserving research data

- Ensure that research results are verifiable
- Make it possible to re-use and build on existing data

Illustration: Gerd Altmann, Pixabay (CC BY)

dms@slu.se

Publishing research data

- Authorities
- Funders
- Publishers
- Universities

Illustration: Gerd Altmann, Pixabay (CC BY)

Better research

- Reduced duplication of effort, accelerated science and innovation
- Research integrity
- Democracy and transparency
- Impact and recognition
- New collaborations

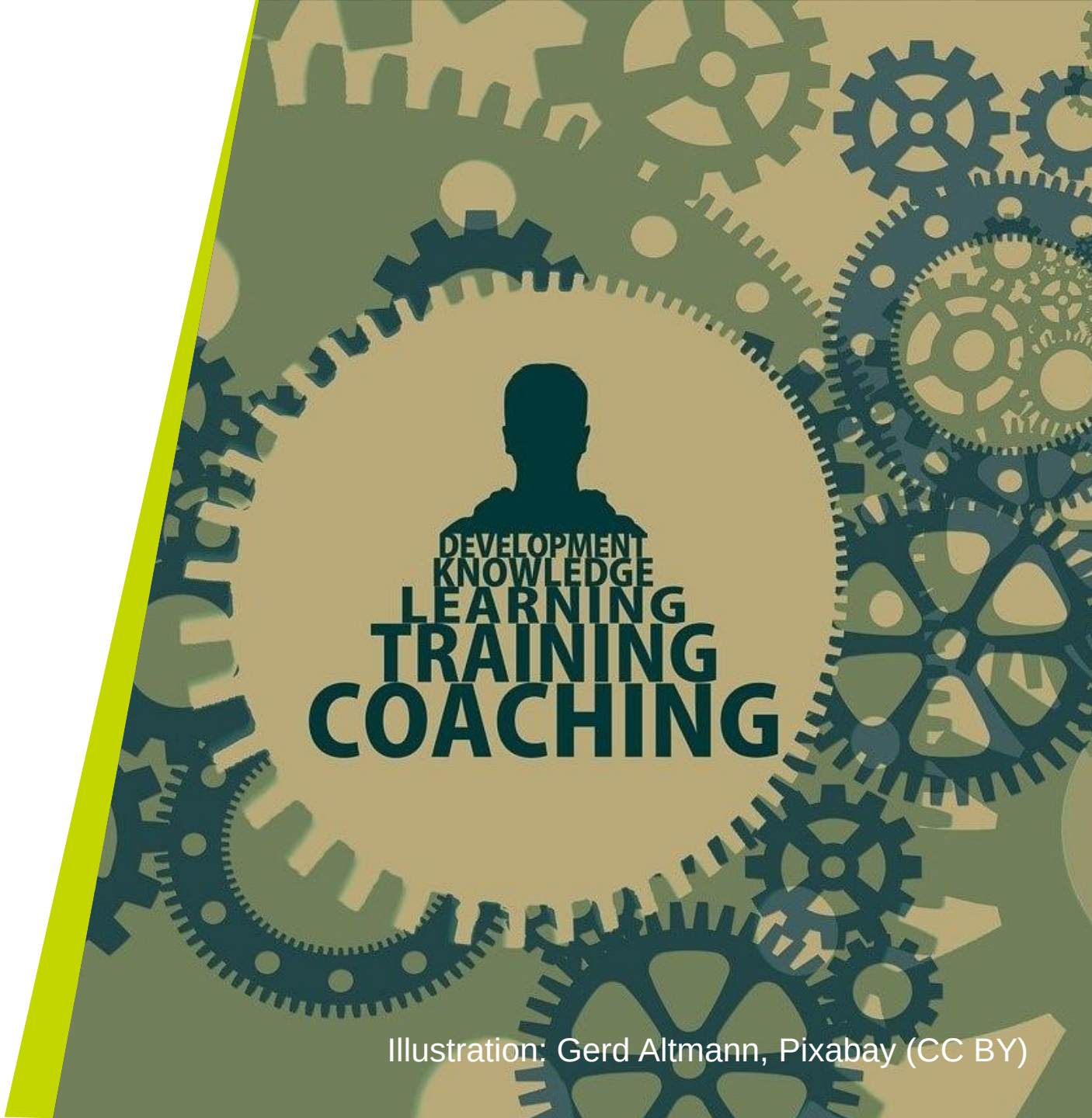
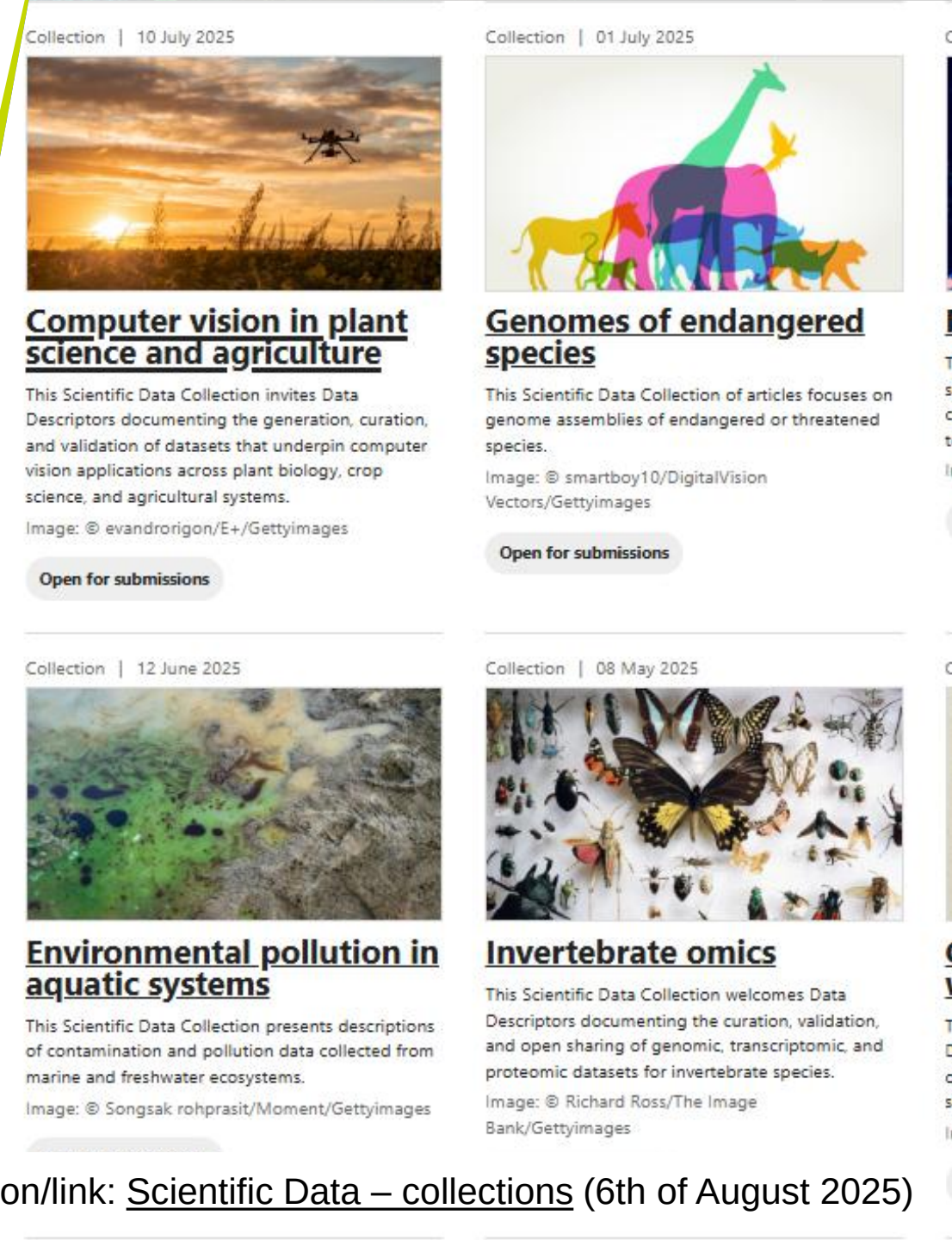


Illustration: Gerd Altmann, Pixabay (CC BY)

Where and how to publish data?

- Data availability statements
- Supplementary material
- Data repositories
- Data papers

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search... Search



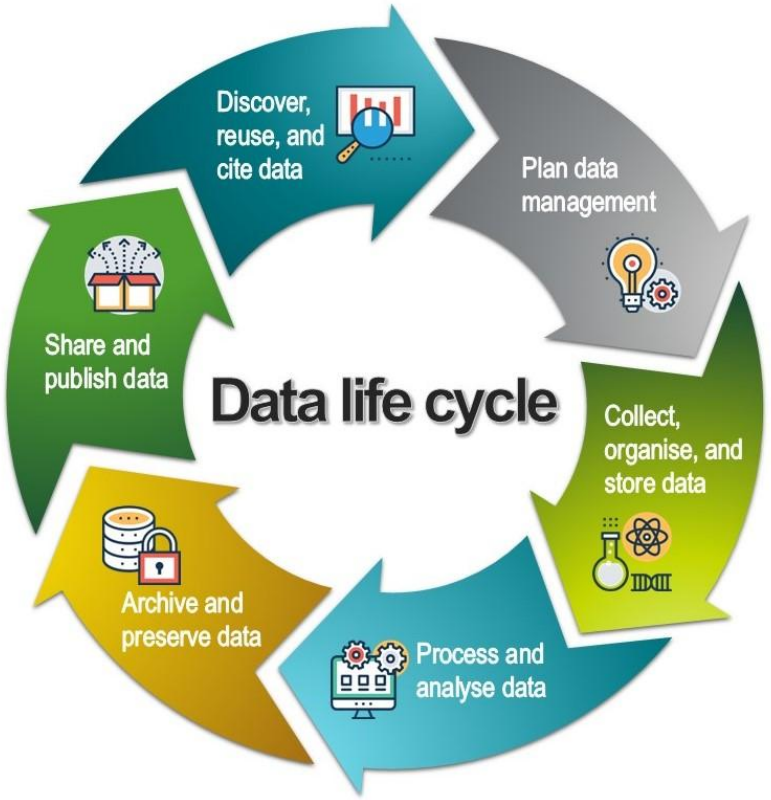
Illustration/link: Scientific Data – collections (6th of August 2025)

DATA
MANAGEMENT
SUPPORT

Research data management: Best practice

[Data Sharing and Management Snafu in 3 Short Acts]

Data management is key



Open data → FAIR data → Data quality
Data management

What is data quality?

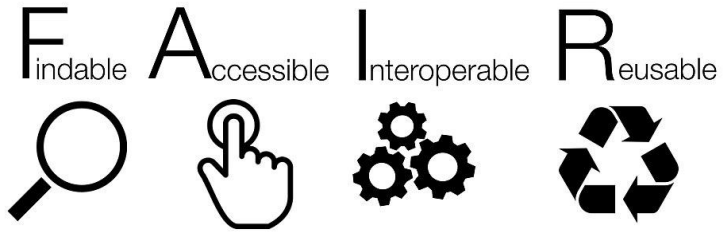
- Accuracy (from verifiable sources)
- Timeliness (available at needed time)
- Validity (within specified domains)
- Consistency (consistent, no duplicates)
- Integrity (relations within tables consistent)
- Completeness (all necessary data included)

Illustration: Gerd Altmann, Pixabay (CC BY)

Open data vs. FAIR data

- “Open data and content can be freely used, modified and shared by anyone for any purpose.”
  Open Knowledge: the Open definition
- Principle for publicly funded research: “as open as possible, as closed as necessary”
  Council of the European Union: conclusions on the transition to an open science system (2016)
- “Data can and should be FAIR even when access is restricted.”
  Horizon Europe: Programme Guide (2024)

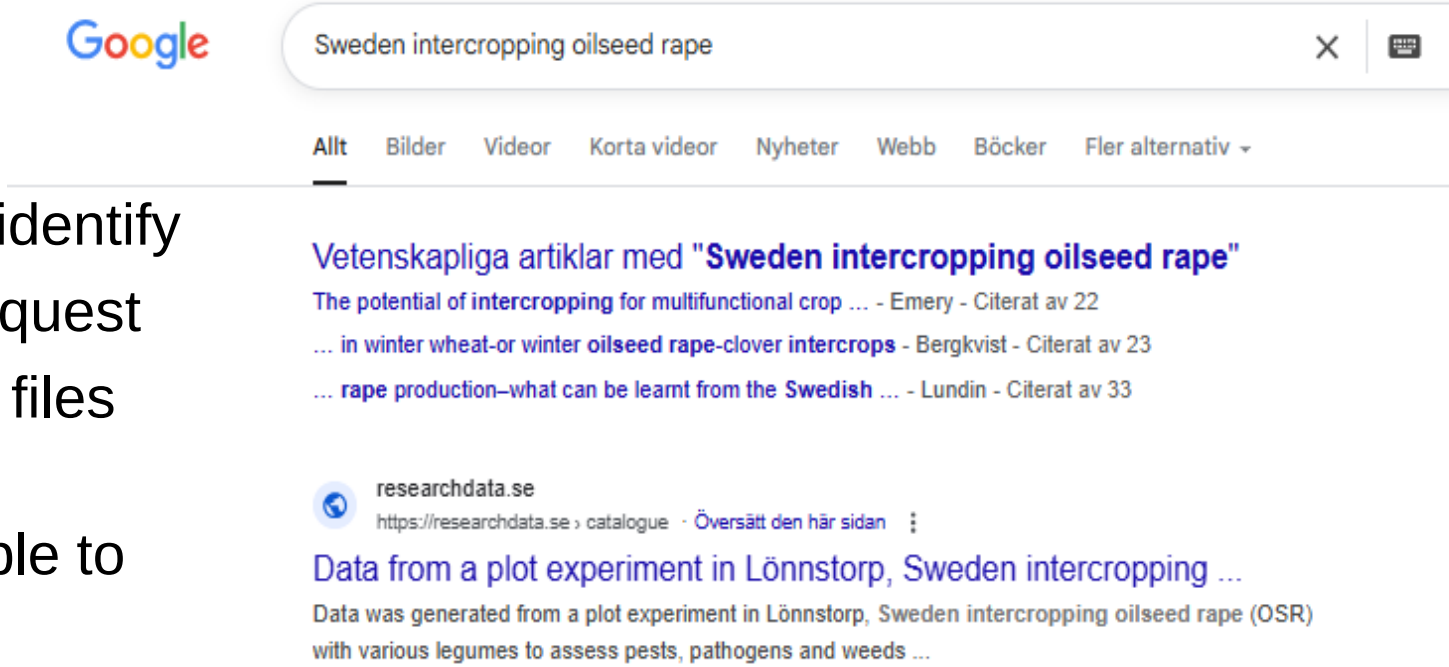
Illustration: Tim Mossholder, Unsplash (CC BY)



FAIR principles

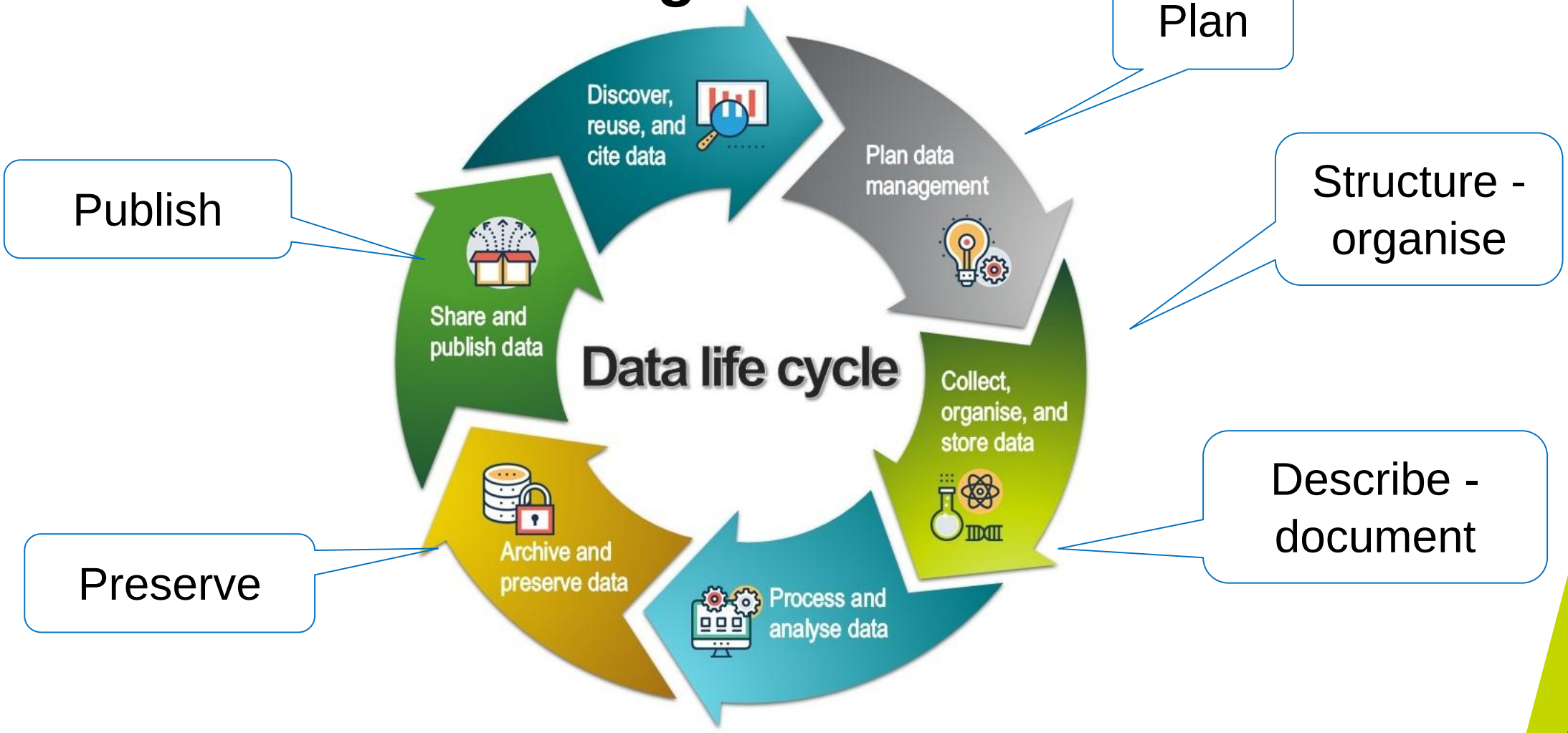
Data should be

- Findable: possible to discover and identify
- Access: possible to download or request
- Interoperable: possible to open the files and understand the contents
- Reuse: permitted to use and possible to cite



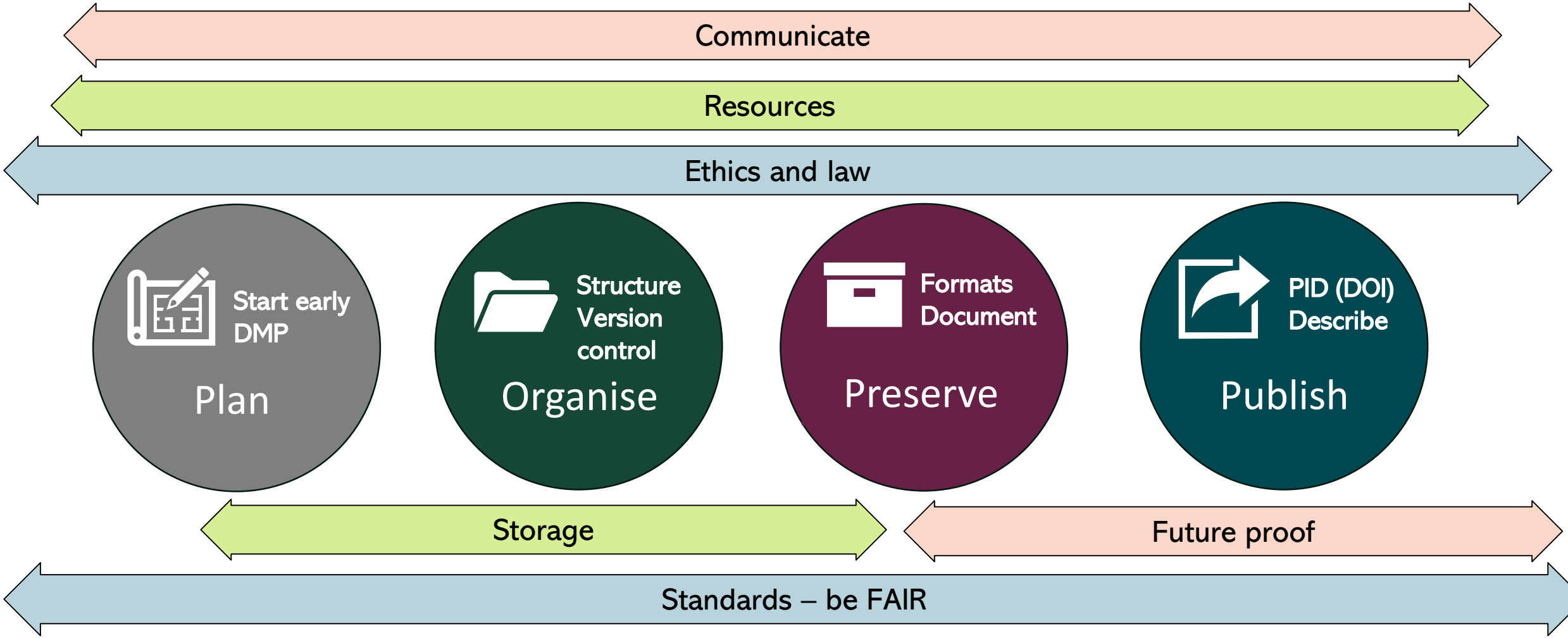
Data from a plot experiment in Lönnstorp, Sweden intercropping oilseed rape (OSR) with various legumes to assess pests, pathogens and weeds 2019-2020, https://doi.org/10.5878/ppwm-4826

Research data management



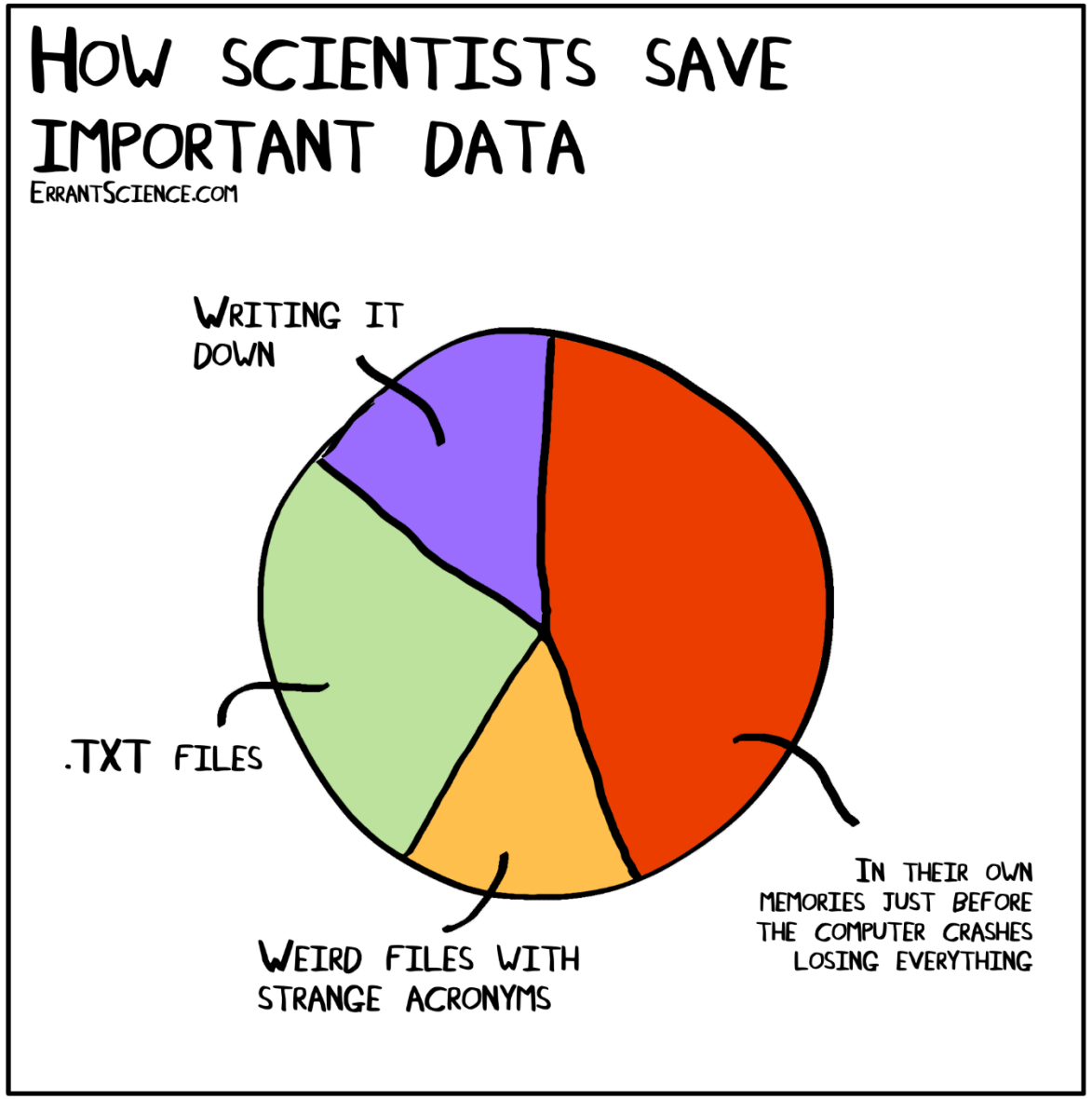
Kanza & Knight: Behind every great research project is great data management

Top 10 tips



Storage

- Information security
- Backup
- Separate raw and active data



Cartoon by errantscience.com

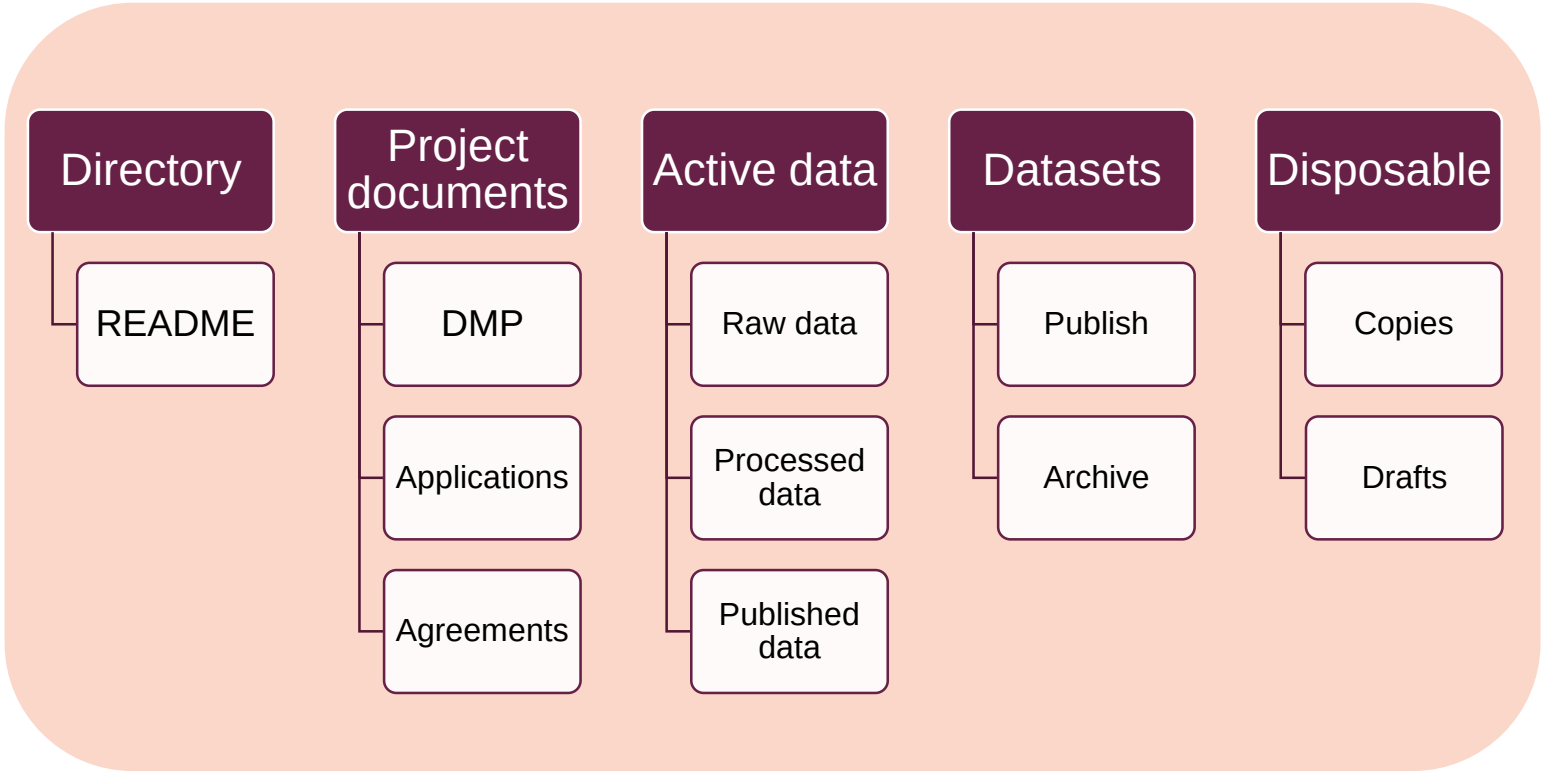
Organisation – structure

Keep the workspace tidy

- Set up a folder structure (directory)
- Not too many top-level folders
- Organise data logically
- Decide how to name folders and files
- Stick to the routines
- README file

Keep track with documentation

- Version control
- Log/electronic notebook



Formats

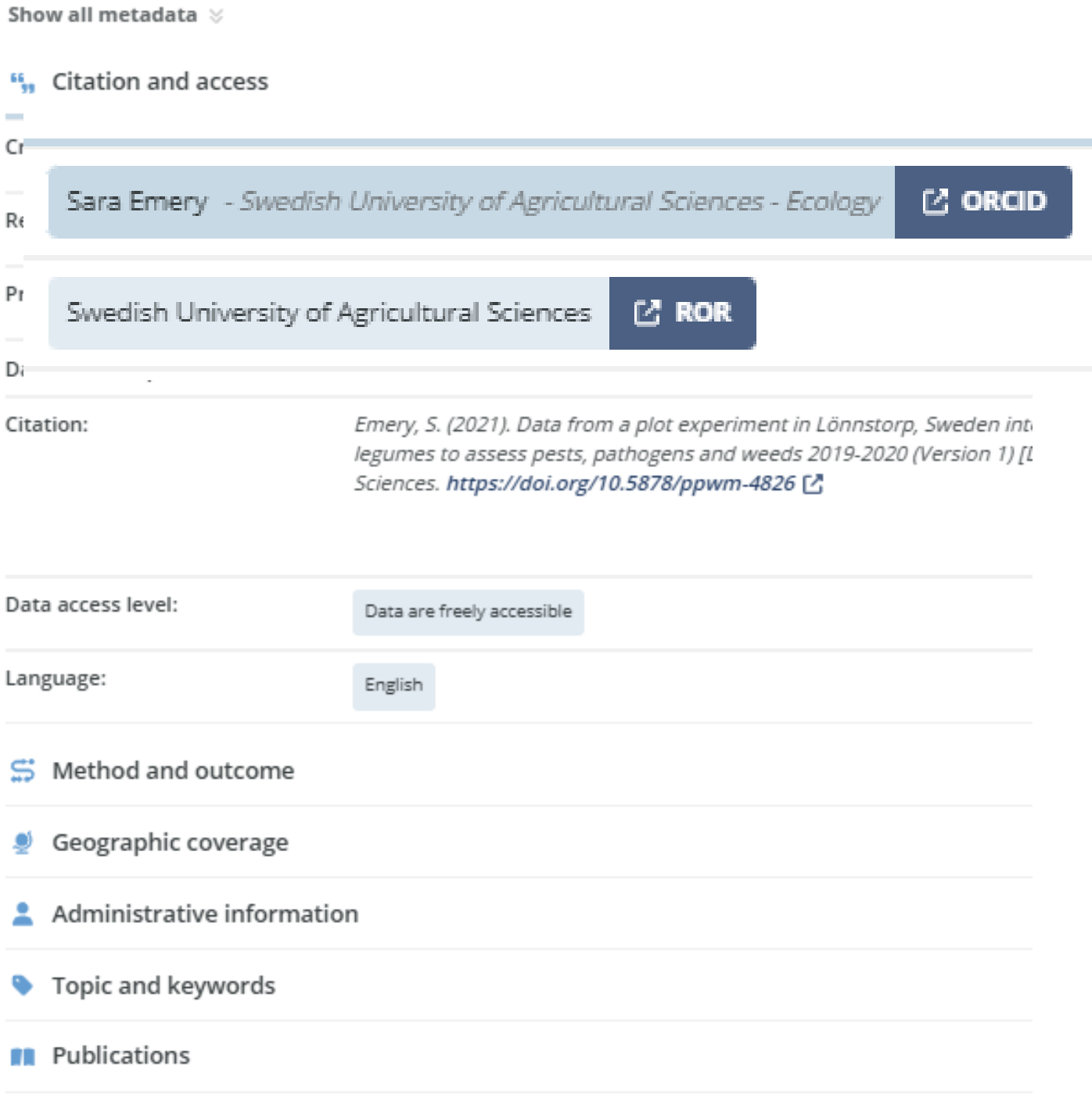
- Formats recommended for text, spreadsheets, audio, video, etc.
- Open, commonly used and machine-readable file formats that can be opened with standard or open-source software.
- Formats that ensure accessibility and long-term preservation

Illustration: Miguel Angel, Vecteezy (CC BY)

Documentation and metadata

- Project level
- Dataset level
- Variable level

- Standards, established terminology



Data from a plot experiment in Lönnstorp, Sweden intercropping oilseed rape (OSR) with various legumes to assess pests, pathogens and weeds 2019-2020, https://doi.org/10.5878/ppwm-4826

Dataset level

Sampling takes place across 4 blocks in random block arrangement with 1 treatment plot per block replicate. Treatment, TreatNum, Block and BlockID are given for each plot of each treatments sampled.

OSR = Oilseed rape
Sfaba = Spring faba

Grams of Oilseed rape

Grams of Weeds in the OSR row

Grams of Weeds between the OSR rows

Total grams of Weeds

Treatment	TreatNum	Block	BlockID	OSR.g	WeedsInRow.g	WeedsBtwnRow.g	WeedsTotal.g
OSR	1	1	1	472.6	68	12.5	80.5
OSR	1	2	14	354	103	33.5	136.5
OSR	1	3	16	438	128.5	25	153.5
OSR	1	4	25	464	54.5	11.5	66
OSR+Sfaba	*2*	*1*	*5*	*321.5*	*47.5*	*19*	*66.5*

CSV files - comma separated values

RowID or PostID

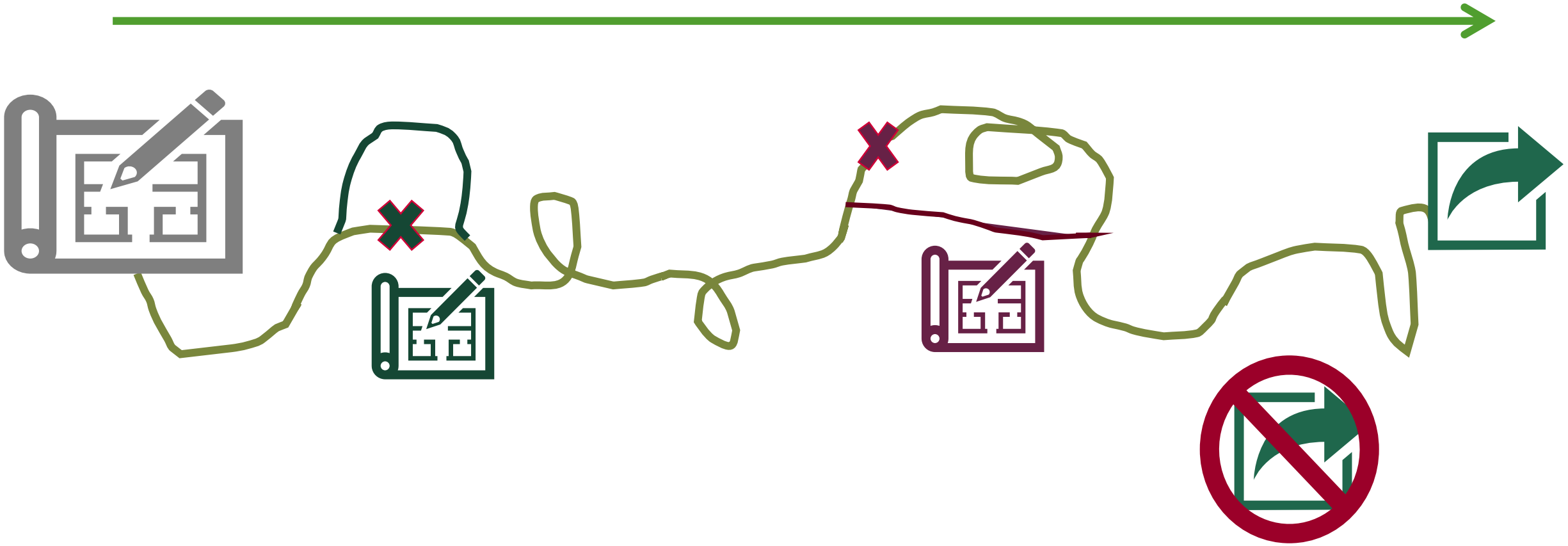
Date and time – one format

No empty cells

No graphic markers

dms@slu.se

Plan ahead



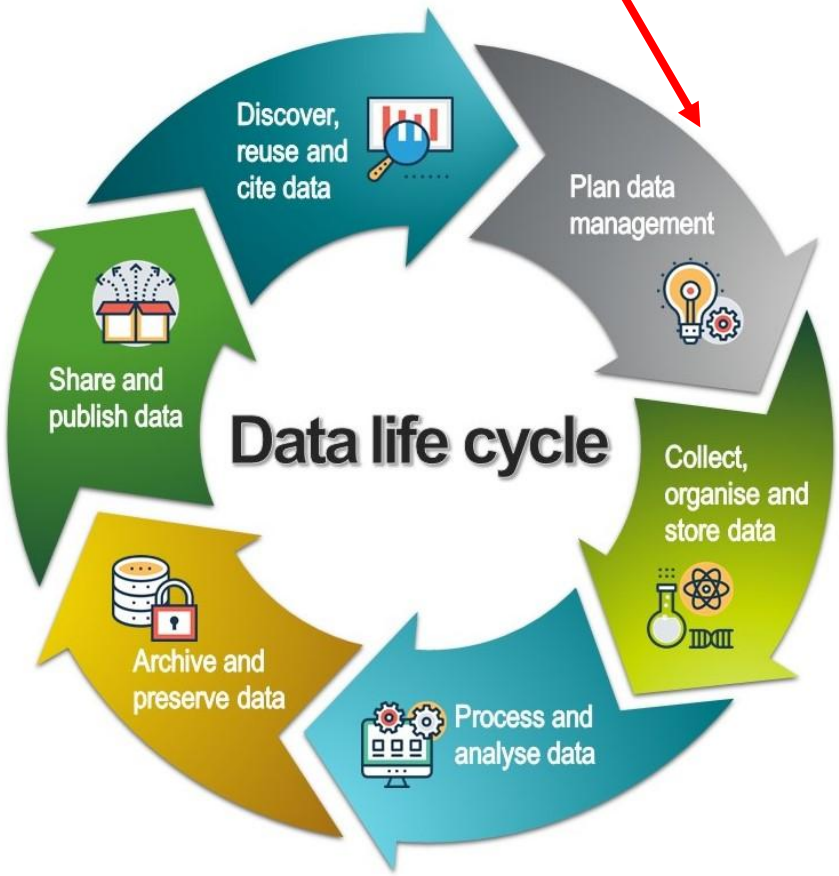
Data Management Plan (DMP)

1. Description of data, e.g.
   - What types, formats and amounts of data do you plan to collect/generate?
   - How will they be collected/generated?
2. Documentation and data quality, e.g.
   - How will data be structured and described to be understandable and (re)usable?
   - What standard terminologies will be used?
   - How will data quality be safeguarded
3. Storage and backup, e.g.
   - How will data be stored and backed up to avoid data loss?
   - How will sensitive and personal data be handled in a controlled and secure way?
4. Legal and ethical aspects, e.g.
   - What ethical and legal issues will need to be considered?
   - How will you make sure data are handled according to legal and ethical requirements?
5. Accessibility and long-term storage, e.g.
   - How, when and where will research data be made discoverable and accessible?
   - How will long-term preservation be safeguarded?
6. Responsibility and resources, e.g.
   - Who is responsible for data management?
   - Who will ensure the resources?

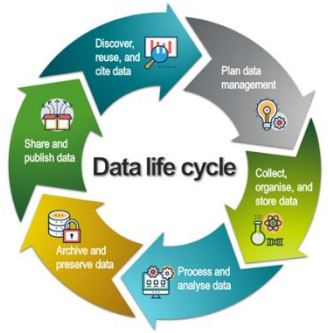
Data management plan

Data management plan (DMP) – a good place to start

- Description of how you plan to handle data throughout a project.
  - What kind of data will I generate?
  - How much data will the project produce?
  - Where will I store data?
  - Does data contain any sensitive information that I can't share?
  - Where will I publish data?

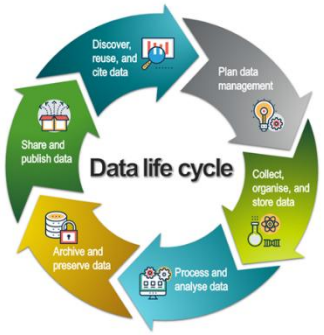


Plan data management



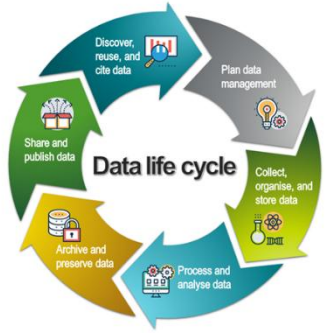
- Address data management in proposals/grant applications/consider your funder's data management and data publishing requirements
- Consider legal and ethical aspects that may affect data management (ethical review, consents, GDPR etc.).
- In collaboration research activities, make agreements regarding responsibilities for data (storage, long term preservation, publishing etc.)
- Use a data management plan (DMP)
- Contact the data management support function

Collect, organise and store data



- Use University storage that is covered by backup and security routines
- Choose open (non proprietary), widely used and machine readable file formats
- Apply a logical and consistent system for organising and naming files
- Format, label and describe data, using common standards and terminologies/controlled vocabularies
- Maintain documentation on a study level (context, methods etc.), file level and data element level (variables, allowable values, abbreviations etc.)

Collect, organise –Start Early!



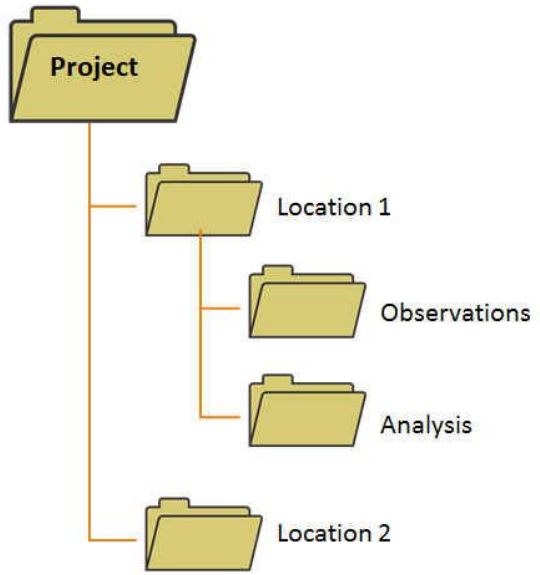
Apply a logical and consistent system for organising and naming files

- FOLDERS
  - Before you start, decide very early on how you will organize folders
  - This helps keep your data organized from the onset
  - Project Name-, type of data you will collect

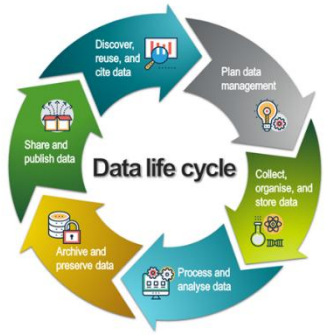
https://www.youtube.com/watch?v=3MEJ38BO6Mo

https://libguides.princeton.edu/c.php?g=102546&p=930626

Example



Collect, organise –start at early

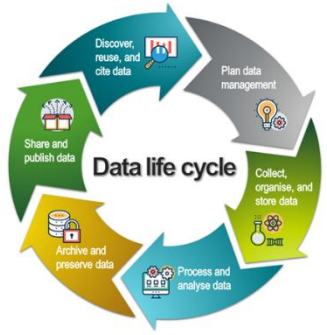


Naming Files and Folders

- Four rules:
  - Be consistent!
  - Choose good names
  - Year format YYYY-MM-DD
  - Create a Data Dictionary

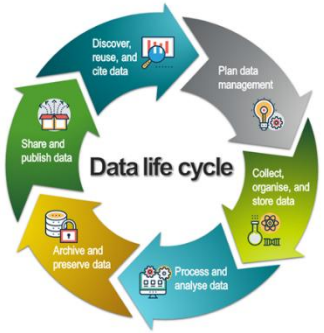
https://www.youtube.com/watch?v=gYDb-GP1CA4
It is important to think about naming conventions of files and folders **BEFORE** you start data collection
Ensures consistency and continuity in recordkeeping
https://guides.lib.purdue.edu/c.php?g=353013&p=2378293
https://libraries.mit.edu/data-management/plan/write/

Collect, organise –start at early



- During the creation phase- collect data
- Put it in spreadsheets:
- Spreadsheets are often used as a multipurpose tool
  - data entry, storage, ~~analysis, and visualization~~
- spreadsheets are **best suited** to data entry and storage

Collect, organise –start at early



- Always organize spreadsheet data in away that both humans and computer programs can read
  - are less error-prone, easier for computers to process, and
  - **easier to share** with collaborators and the public
    - https://datadryad.org/stash
- It has become common now that publishers request for your data to publish along with your work

Generate a raw data file

- Here raw data means data that has not yet been processed
- The best format for a raw data file is the “rectangular format”. See Browman & Woo, 2018

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY

Received June 2017
Revised August 2017

KEYWORDS

Data management; Data organization; Microsoft Excel; Spreadsheets

Icon by Freepik, Flaticon

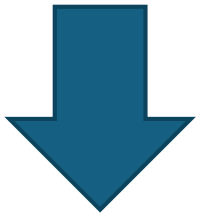
Make spreadsheet a rectangle

A

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102		95.3
4	103	2015–06–18	97.5
5	104		117.0
6	105		108.0
7	106	2015–06–20	149.0
8	107		169.4

B

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334

Icon by Freepik, Flaticon

Make spreadsheet a rectangle

	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
4	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
5	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
6	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
7	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

	A	B	C	D	E	F
1	mouse_id	sex	week	date	glucose	ght
2	3005	M	4	3/30/2007	19.3	635
3	3005	M	6	4/11/2007	31	460.7
4	3005	M	8	4/27/2007	39.6	530.2
5	3017	M	4	10/6/2006	25.9	202.4
6	3017	M	6	10/19/2006	45.1	384.7
7	3017	M	8	11/3/2006	57.2	458.7
8	3434	F	4	11/22/2006	26.6	238.9

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

Naming conventions

- Make sure you name the variables in a logical way that is easy to remember
- Variable names should be short, no spaces no special characters- see Browman & Woo, 2018

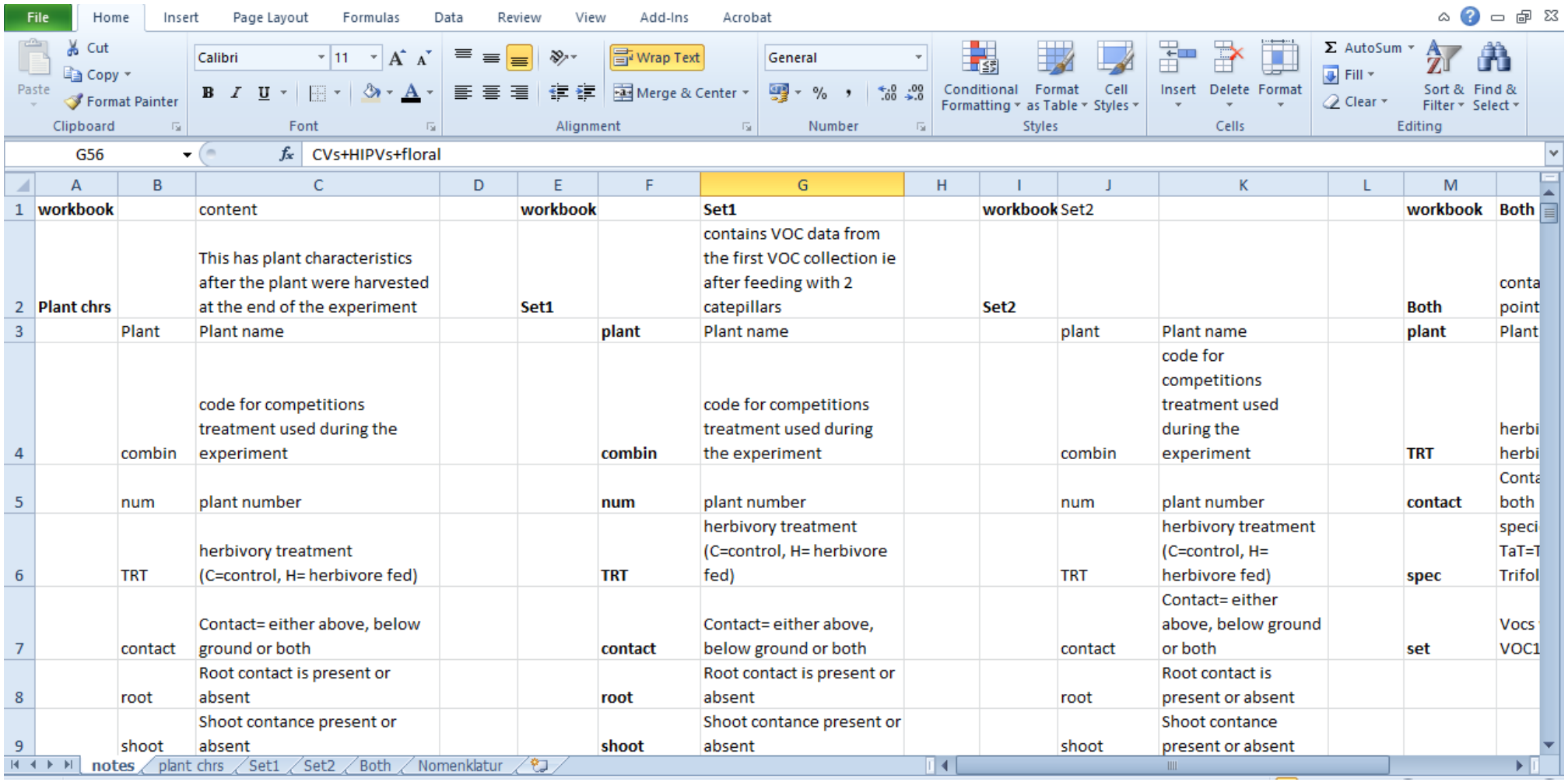
Table 1. Examples of good and bad variable names.

good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

Icon by Freepik, Flaticon

Data Dictionary

- Your raw data file must have a data dictionary Browman & Woo, 2018



	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	**workbook**		content		**workbook**		**Set1**		**workbook**	Set2			**workbook**	**Both**
2	**Plant chrs**		This has plant characteristics after the plant were harvested at the end of the experiment		**Set1**		contains VOC data from the first VOC collection ie after feeding with 2 catepillars		**Set2**				**Both**	conta point
3		Plant	Plant name			**plant**	Plant name			plant	Plant name		**plant**	Plant
4		combin	code for competitions treatment used during the experiment			**combin**	code for competitions treatment used during the experiment			combin	code for competitions treatment used during the experiment		**TRT**	herbi herbi
5		num	plant number			**num**	plant number			num	plant number		**contact**	Conta both
6		TRT	herbivory treatment (C=control, H= herbivore fed)			**TRT**	herbivory treatment (C=control, H= herbivore fed)			TRT	herbivory treatment (C=control, H= herbivore fed)		**spec**	speci TaT=T Trifol
7		contact	Contact= either above, below ground or both			**contact**	Contact= either above, below ground or both			contact	Contact= either above, below ground or both		**set**	Vocs VOC1
8		root	Root contact is present or absent			**root**	Root contact is present or absent			root	Root contact is present or absent			
9		shoot	Shoot contance present or absent			**shoot**	Shoot contance present or absent			shoot	Shoot contance present or absent			

Data Dictionary

- Variable= **Borehole**
- Explanation= Community management of access to water from boreholes
- Variable= **Ponds**
- Explanation= Seasonal management of XX
- Code book= Data Dictionary

Exercise

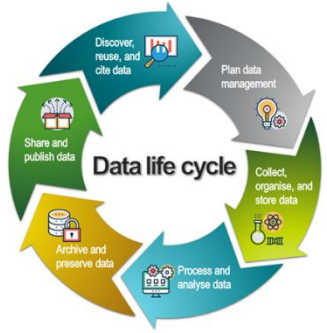
- A study was conducted to test the question as to whether ***levels of iron*** differed with ***increase in depth*** at Kilifi bay.
- Water samples were taken at two depths and the level of iron measured in ng/liter. The results are displayed in the table below.

Depth	rep1	Rep2	Rep3	Rep4
0 cm	5	4	6	6
30 cm	7	6	5	6

- Create a “rectangle” file with the data above
- Save the file as CSV format
- Import the .csv to R-cloud
- Save the file in the data folder that you created

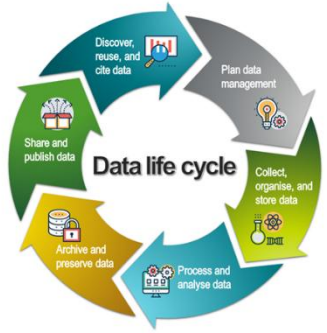
Icon by Freepik, Flaticon

Process and analyse data



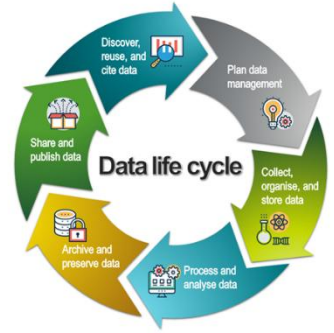
- Make a working-copy of collected data for processing and analysis, keeping original raw data intact
- Document data actions and steps in data analysis (cleaning, validation, aggregation, coding, etc.)
- Save and backup iterations of data and apply a clear and consistent **versioning system**
- Document and preserve code and scripts
- If specialised software is used for data processing or analysis, export data to an open format

Archive and preserve data



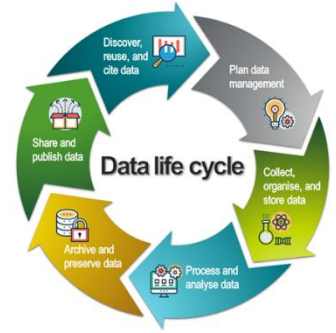
- Save data in open and widely used file formats
- Keep data files in good order and be prepared that access to data can be requested during data collection and before the project is finalised
- Do not disperse, remove or dispose of research data without a prior legal decision to do so
- Secure the future of data by archiving them – in safe institutional servers/repositories!
- Archive data with associated documentation to enable interpretation and reuse in the future

Share and publish data



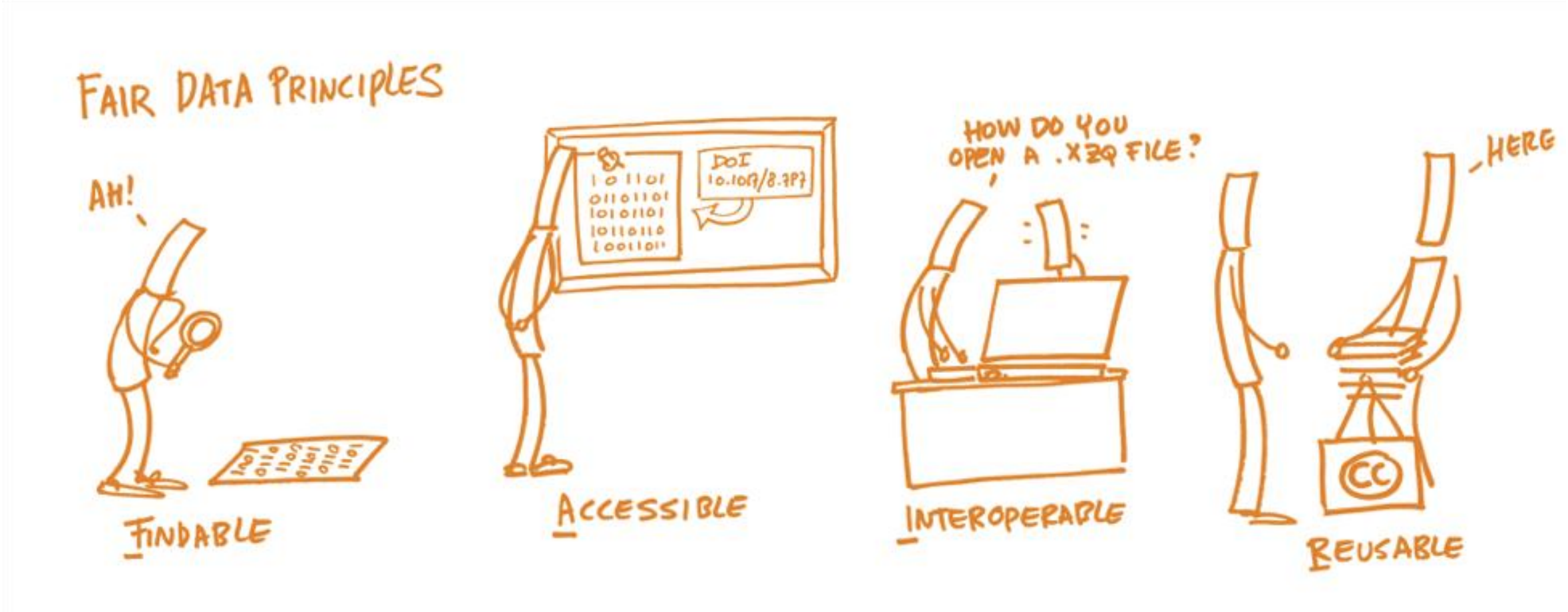
- At an early stage consider , what can and will be shared, how and under what conditions
- Publish data in a data repository under the principle “as open as possible, as closed as necessary”, and receive a persistent identifier (e.g. DOI)
- Include documentation that describes data files and data variables
- Also publish questionnaires, protocols, code etc. that may facilitate data interpretation, reproducibility and reuse
- Add a data availability statement in your article and cross link between article and dataset

Discover, reuse and cite data

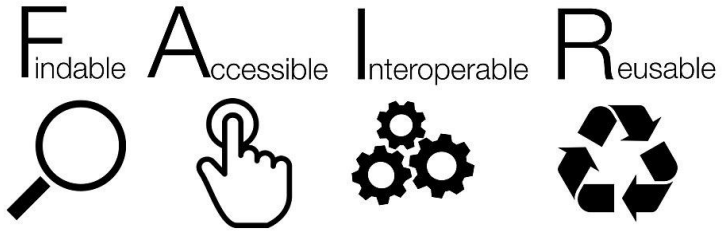


- Before deciding to collect new data, investigate if there are existing data that could be used to answer your research question
- Assess reliability and quality
- Check terms of use and obtain permissions if required (legal/ethical)
- Account for the origin/location of data and how they were applied in your study
- Cite properly

Data Management Plan- FAIR principles



F- Fair



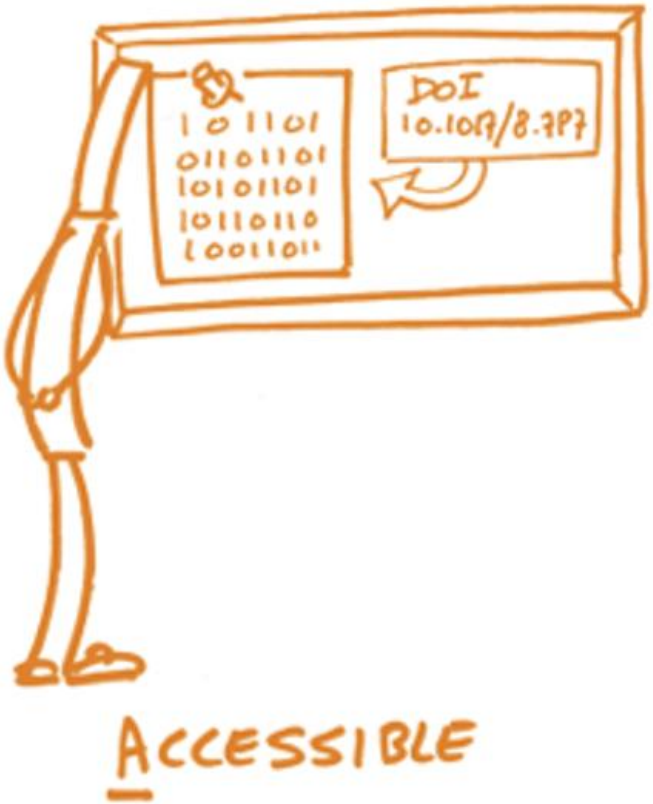
- To be **Findable** any Data Object should be uniquely and persistently identifiable [4]

  1.1. The same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata, [4 and JDDCP 4 and JDDCP 6]

  1.2. A Data Object should minimally contain basic machine **actionable metadata** that allows it to be distinguished from other Data Objects [see JDDCP 5]

  1.3. Identifiers for any concept used in Data Objects should therefore be **Unique** and **Persistent** [5 and JDDCP 4 and JDDCP 6].

A- Accessible



2. **Data is Accessible** in that it can be always obtained by machines and humans

2.1 Upon appropriate authorization authorization [6]

2.2 Through a well-defined protocol [7 and JDDCP 5]

2.3 Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object

I- Interoperable

3. Data Objects can be **Interoperable** only if:

3.1. (Meta) data is machine-actionable [8]

3.2. (Meta) data formats utilize shared vocabularies and/or ontologies [9]

3.3 (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible [10]

R- Re-usable

4. For Data Objects to be **Re-usable** additional criteria are:

4.1 Data Objects should be compliant with principles 1-3

4.2 (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources [11 and JDDCP 7 and JDDCP 8]

4.3 Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation (ref to JDDCP 1-3).