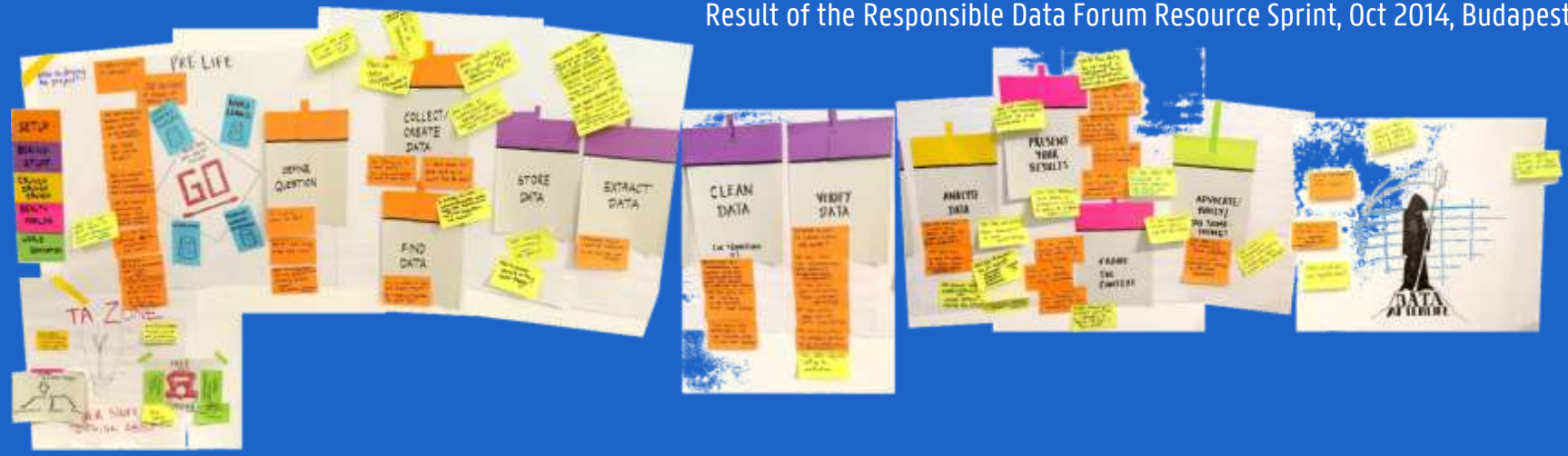


The data lifecycle.
Result of the Responsible Data Forum Resource Sprint, Oct 2014, Budapest. Wikipedia, CC-BY-SA

DATA MANAGEMENT PLANNING FOR EFFICIENT RESEARCH

Laura Standaert, data steward Life Sciences Ghent University

GHENT UNIVERSITY

WHAT COULD GO WRONG?

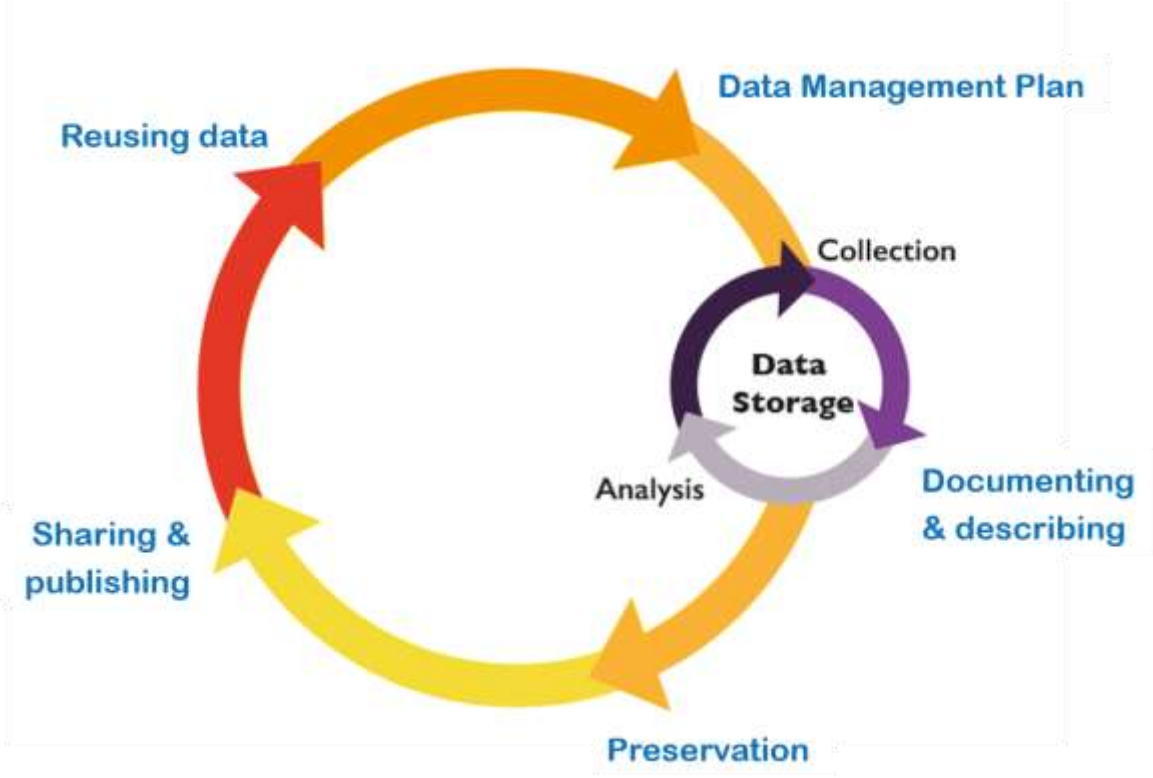
- 3 years into the project, data is lost
- Data exists but can't be reused (no documentation, corrupted, outdated format…)
- You are required to share data but you didn't ask for consent
- Disagreement between project partners on exploitation of results
- Can't keep track of your own files

GHENT UNIVERSITY

DATA MANAGEMENT PLAN (DMP)

- Outlines how research data will be handled **during** & **after** a project.
- Increasingly required by research funders/institutions

  -> as a **first step** towards good **research data management** (RDM)



Adapted from DCC

GHENT UNIVERSITY

MAIN DMP SECTIONS

Data summary / description

Purpose, origin, type, format & volume of the data

Ethical and legal issues

IRB approval, IP, GDPR

Documentation and metadata

Where and how, standards, ontologies, vocabularies

Resources

Costs

Responsibilities

Owner and roles

Data sharing and reuse

Repositories, terms, contracts, licenses

Data storage during the project

Security, back-up, transfer & access

Long-term preservation

What, where, restrictions?

GHENT UNIVERSITY

DMP: FIRST VERSION VS. LAST VERSION

Data Management Plan
First submission

- First scope of reusability.
- Basic information about your data.
- Data storage and back up vs. long term preservation.
- Initial idea about metadata standards and dedicated archive / repository.
- Identification of legal and/or ethical issues.
- Establishing responsibilities.

Data Management Plan
Final submission

- Definite selection of reusable outputs.
- Precise description of all data types.
- Clear long-term preservation strategy: selected data archive(s) and/or repository(ies).
- FAIR data and open data ("As open as possible, as closed as necessary").

GHENT UNIVERSITY

Adapted from: Jasmin Böhmer. (2019, September). UBC RDM Workshop 1 - Data Management Planning (Version V2). Zenodo. http://doi.org/10.5281/zenodo.3469953

DMP SOFTWARE

DMPONLINE.BE

GHENT
UNIVERSITY

DMPONLINE.BE TUTORIAL

https://github.com/DMPbelgium/Guidance/wiki/End-user-manual

GHENT
UNIVERSITY

https://www.youtube.com/watch?v=5U9_DDdBzYk

WHAT TO COVER IN YOUR DMP

GHENT UNIVERSITY

MAIN DMP SECTIONS

Data summary / description
Purpose, origin, type, format & volume of the data

Ethical and legal issues
IRB approval, IP, GDPR

Documentation and metadata
Where and how, standards, ontologies, vocabularies

Resources
Costs

Responsibilities
Owner and roles

Data sharing and reuse
Repositories, terms, contracts, licenses

Data storage during the project
Security, back-up, transfer & access

Long-term preservation
What, where, restrictions?

GHENT UNIVERSITY

RESEARCH DATA SUMMARY

- List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project.
- If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type

GHENT UNIVERSITY

RESEARCH DATA IN RDM

"**any information** that has been collected, observed, generated or created **to validate original research findings**. "

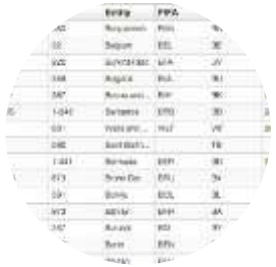
Leeds University

- **Role** in the research project rather than nature
- **Foundations of your scientific claims** rather means to and end

Survey

Measurement

Observation

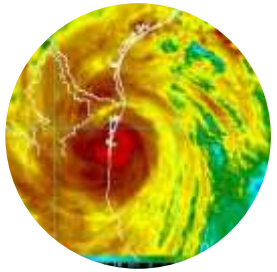


Code

Image

Document

Model



GHENT UNIVERSITY

DATA DESCRIPTION – SOME TIPS

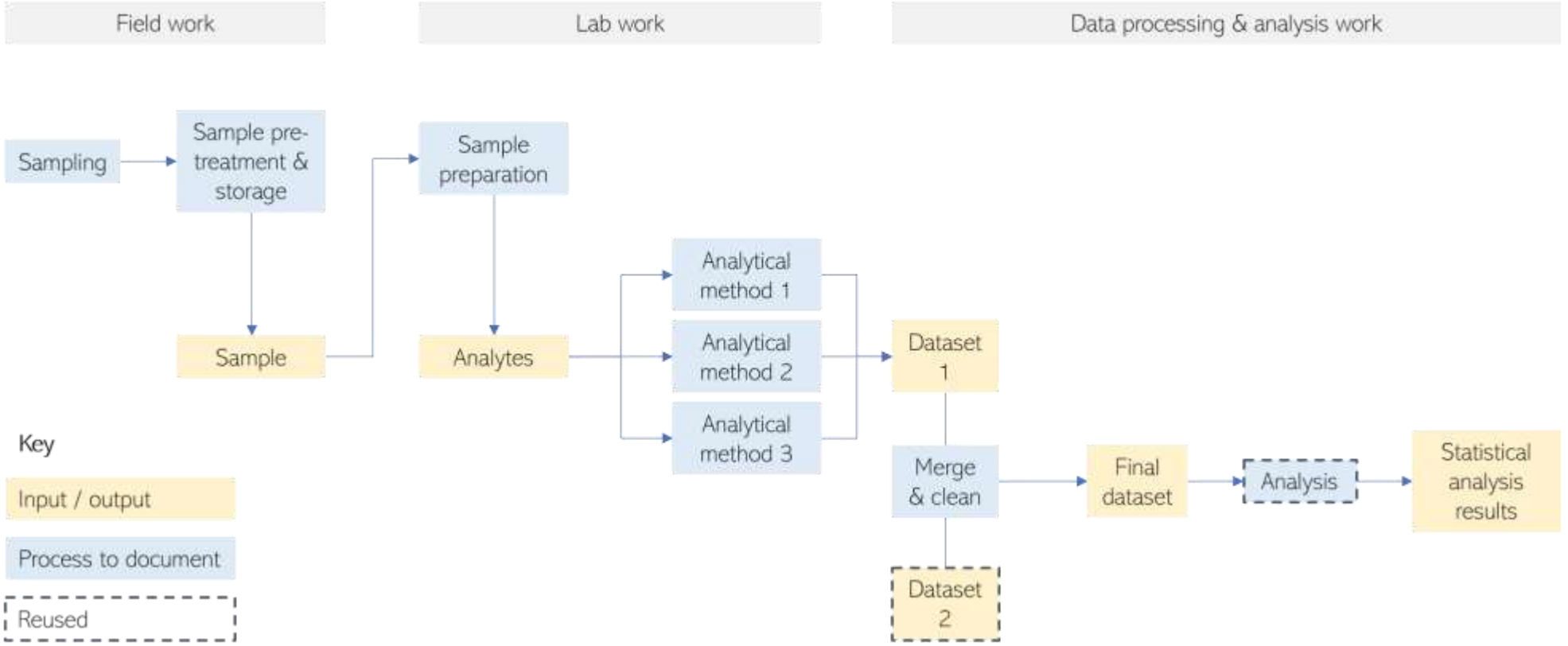
- Only provide answers relating to research data, not publications!
- Remember that 'research data' come in many different forms
- Clear distinction between re-used and new data.
- Break down and list your data types conveniently: e.g. by technique, by purpose, by research phase.
- Provide enough details for outsiders to understand the sort of data involved
  - e.g. distinguish between digital/non-digital data, quantitative/qualitative data, raw & processed data
  - specify file formats (preferably no proprietary/unusual formats)
  - specify data collection methods (e.g. experimental, observational, simulation… data)

DATA DESCRIPTION: PREPARATION

- Complex project structure, identify datatypes

⇒ Design a data flow

EXAMPLE



GHENT UNIVERSITY

DATA DESCRIPTION: DATA TABLE APPROACH

EXAMPLE

Dataset	Description	Data type	Format	Expected Volume
Farmer-survey	Questionnaire for farmers in Flanders	Observational qualitative	Paper and pencil + digitized transcript (.docx)	<1 GB, paper surveys of 300 participants
			Smartphone based questionnaire. Exported tables as .csv	<1 GB
Tracker data	Activity tracker for data subjects	Observational quantitative	Physical activity and sedentary time in .GT3X files	250 KB per subject x 300 subjects = 75 MB
Farmer-interview	Interview with a selection of farmers	Observational qualitative	.MP3 + transcript in .docx	About 5 GB
mRT-activity	Micro-randomized trial to determine the effect of isolated determinants on **physical activity**	Experimental	Digital textual and numerical data combined in spreadsheets and obtained via tracker based experiment. CSV format and GT3X for raw activity tracker data	<5 MB for the CSV files. Raw accelerometer data (.GT3X files) are 250kB per subject x 40 subjects per experimental study = 10MB
mRT-sendentary	Micro-randomized trial to determine the effect of isolated determinants on **sedentary time**	Experimental	Digital textual and numerical data combined in spreadsheets and obtained via a tracker based experiment. CSV format and GT3X for raw activity tracker data	<5 MB for the CSV files. Raw accelerometer data (.GT3X files) are 250kB per subject x 40 subjects per experimental study = 10MB

GHENT UNIVERSITY

DATA DESCRIPTION: DATA TABLE APPROACH

EXAMPLE

Dataset name	Description	New or reused	Data type	Data format	Volume
WP1 Pollutant concentration before and after treatment	Comparison of the concentration of heavy metals and organic pollutants before and after different thermochemical treatment	New	Experimental, quantitative	xlsx, csv	<5MB
WP1 Nutrient bio-availability	Comparison of nutrient bio-availability after different thermochemical treatments	New	Experimental, quantitative	xlsx, csv	<5MB
WP1 Data from statistical analysis	Results from statistical analysis to determine the influence of the dependent variables (chemical treatment/reactors, temperature, retention time) in pollutant removal and nutrient bio-availability.	New	Analysis script & results, quantitative	R, csv, png	<20MB
WP2 Consumers attitudes - questionnaire responses	Data on attitudes towards consumption of food that has been grown using fertilisers based on recycled human excreta. Online questionnaires will be performed using Qualtrics and will collect: ● demographic data from participants: place of residence, age, gender, education, occupation, annual income. ● respondent's opinions or views about issues in relation to the research topic: environmental awareness questions, attitudes or acceptance of the proposed product, etc.	New	Observational, quantitative & qualitative	csv	<5MB
WP2 Consumer attitudes - analysis	Statistical analysis of survey responses	New	Analysis results, quantitative	sav	<10MB

GHENT UNIVERSITY

DATA TABLE(1)

Bonus: Detect RDM needs

There is not a "one size fits all" approach

How much granularity is needed and how do I break down the data into different rows?

- It depends on your project and a good approach is to categorize the content in a way that proves useful for the management of your data
- E.g. Divide your data according to:
  - content type (quantitative vs. qualitative)
  - data collection method (observational, experimental, simulation)
  - by purpose
  - by format
  - differentiate between own data and third-party data

GHENT UNIVERSITY

DATA TABLE(2)

Bonus: Detect RDM needs

How many columns or characteristics should I provide?

- DMP templates will explicitly require some characteristics, e.g. description, new or reused, data type, format and expected volume
- If possible, add other relevant characteristics that are specific to your data and that have implications for data management. E.g.:
  - temporal and geographical scope of the data
  - data is sensitive or not
  - access restrictions

GHENT UNIVERSITY

DATA OVERVIEW TABLE IN FWO/VLAIO DMP

- Data overview tables are a useful way to index all the data/outputs in the project and refer to them in other parts of DMP

				Only for digital data	*Only for digital data*	*Only for digital data*	*Only for physical data*
Dataset Name	**Description**	**New or reused**	**Digital or Physical**	**Digital Data Type**	**Digital Data format**	**Digital data volume (MB/GB/TB)**	**Physical volume**
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … NA	*Please choose from the following options:* • < 100MB • < 1GB • < 2 TB • … • NA	

ETHICAL & LEGAL ISSUES QUESTIONS (1)

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

GHENT UNIVERSITY

ETHICAL & LEGAL ISSUES QUESTIONS (2)

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

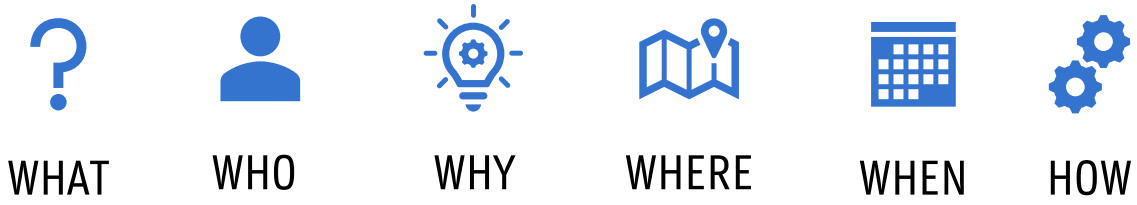
Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use?

GHENT UNIVERSITY

DATA DOCUMENTATION & METADATA QUESTIONS

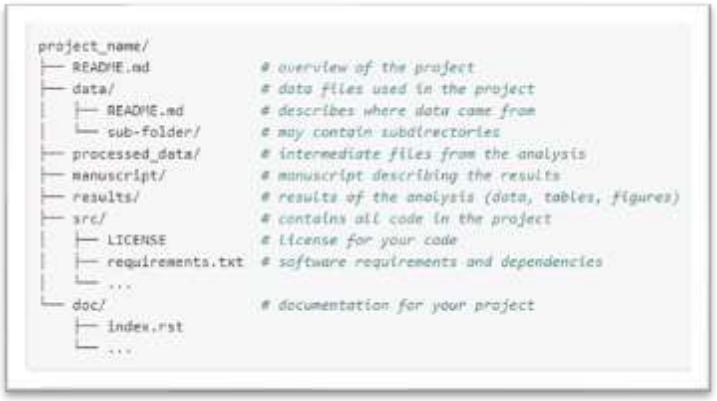
- Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future.
- Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

GHENT UNIVERSITY



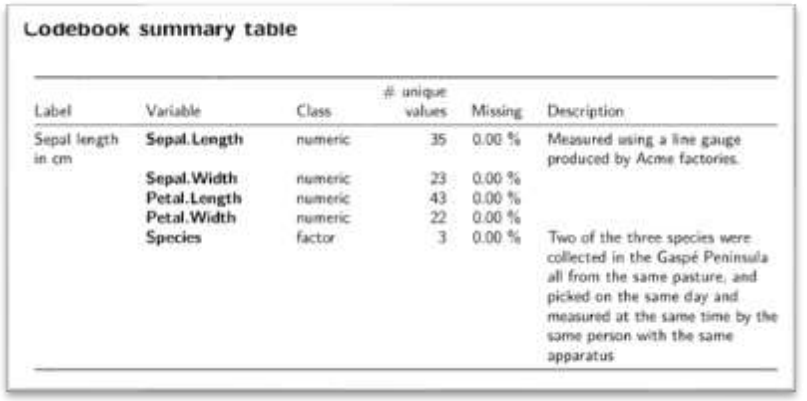
- Review & verification of research
- Reproducibility
- Make data understandable and reusable

GHENT UNIVERSITY



```
project_name/
├── README.md            # overview of the project
├── data/                # data files used in the project
│   ├── README.md        # describes where data came from
│   └── sub-folder/      # may contain subdirectories
├── processed_data/      # intermediate files from the analysis
├── manuscript/          # manuscript describing the results
├── results/             # results of the analysis (data, tables, figures)
├── src/                 # contains all code in the project
│   ├── LICENSE          # license for your code
│   ├── requirements.txt # software requirements and dependencies
│   └── ...
└── doc/                 # documentation for your project
    ├── index.rst
    └── ...
```

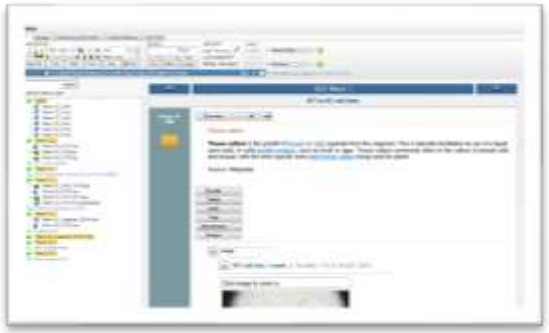
Readme files



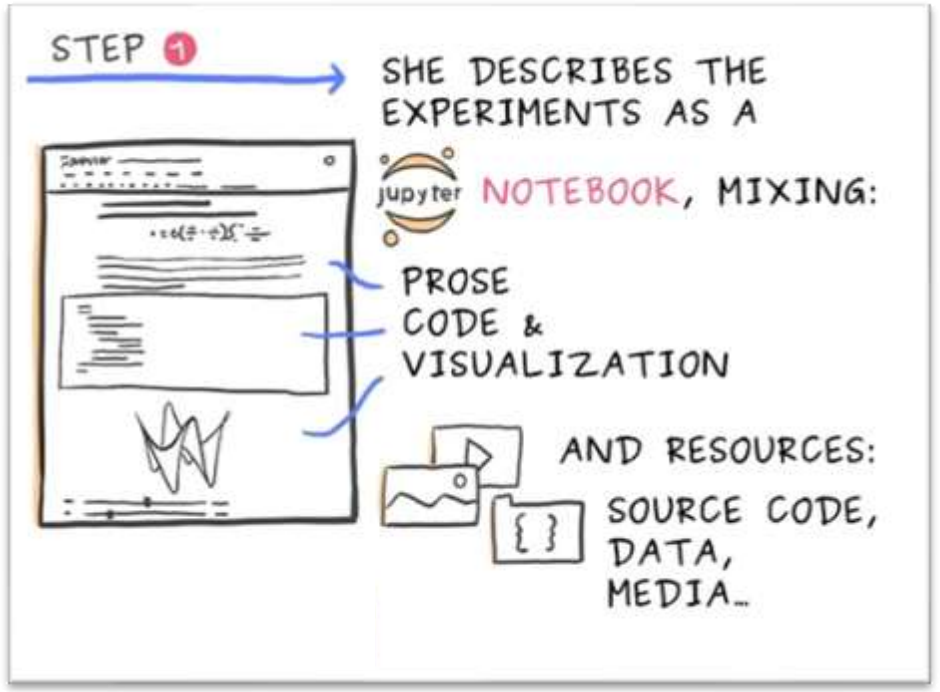
Codebook summary table

Label	Variable	Class	# unique values	Missing	Description
Sepal length in cm	**Sepal.Length**	numeric	35	0.00 %	Measured using a line gauge produced by Acme factories.
	Sepal.Width	numeric	23	0.00 %	
	Petal.Length	numeric	43	0.00 %	
	Petal.Width	numeric	22	0.00 %	
	Species	factor	3	0.00 %	Two of the three species were collected in the Gaspé Peninsula all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus

Codebooks



Paper and electronic lab notebooks



Jupyter notebooks
https://opendreamkit.org/2017/11/02/use-case-publishing-reproducible-notebooks/

GHENT UNIVERSITY

DATA STORAGE & BACKUP QUESTIONS

- Where will data be stored?
- How will the data be backed up?
- Is there sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.
- What are the expected costs for data storage & backup during the project? How will these costs be covered?
- Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorised persons?

GHENT UNIVERSITY

DATA SHARING & REUSE QUESTIONS (1)

- Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.
- If access is restricted, please specify who will be able to access the data and under what conditions.
- Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

GHENT UNIVERSITY

DATA SHARING & REUSE QUESTIONS (2)

- Where will the data be made available? If already known, please provide a repository per dataset or data type.
- When will the data be made available?
- Which data usage licenses are you going to provide? If none, please explain why.
- Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.
- What are the expected costs for data sharing? How will these costs be covered?

GHENT UNIVERSITY

DATA PRESERVATION QUESTIONS

- Which data will be retained for the expected 5-year period after the end of the project? In case some data cannot be preserved, clearly state the reasons for this.
  - Other terms apply to Clinical trials, patient records, 'experiments', …
- Where will these data be archived (= stored and curated for the long term)?
- What are the expected costs for data preservation during the expected retention period? How will the costs be covered?

GHENT UNIVERSITY

RESPONSIBILITIES

QUESTIONS

- Who will manage data documentation and metadata during the research project?
- Who will manage data storage and backup during the research project?
- Who will manage data preservation and sharing?
- Who will update and implement this DMP?

GHENT UNIVERSITY

DMP TIPS

- Be consistent throughout the document.
  - If a dataset cannot be shared, do not mention in a later paragraph that all data will be shared openly on Zenodo
- If certain information is not available yet, mention when and how you will address the issue
- Be complete for every question: mention all datatypes or outputs.
  - E.g.: For WP1 this question is not applicable, for WP 2-3 we will do this, and for WP 4-6 we will do that.

Ghent University Data Stewards

RESEARCH DEPARTMENT - UNIVERSITY LIBRARY

rdm.support@ugent.be
@UGentRDM
ugent.be/en/research/datamanagement

GHENT UNIVERSITY