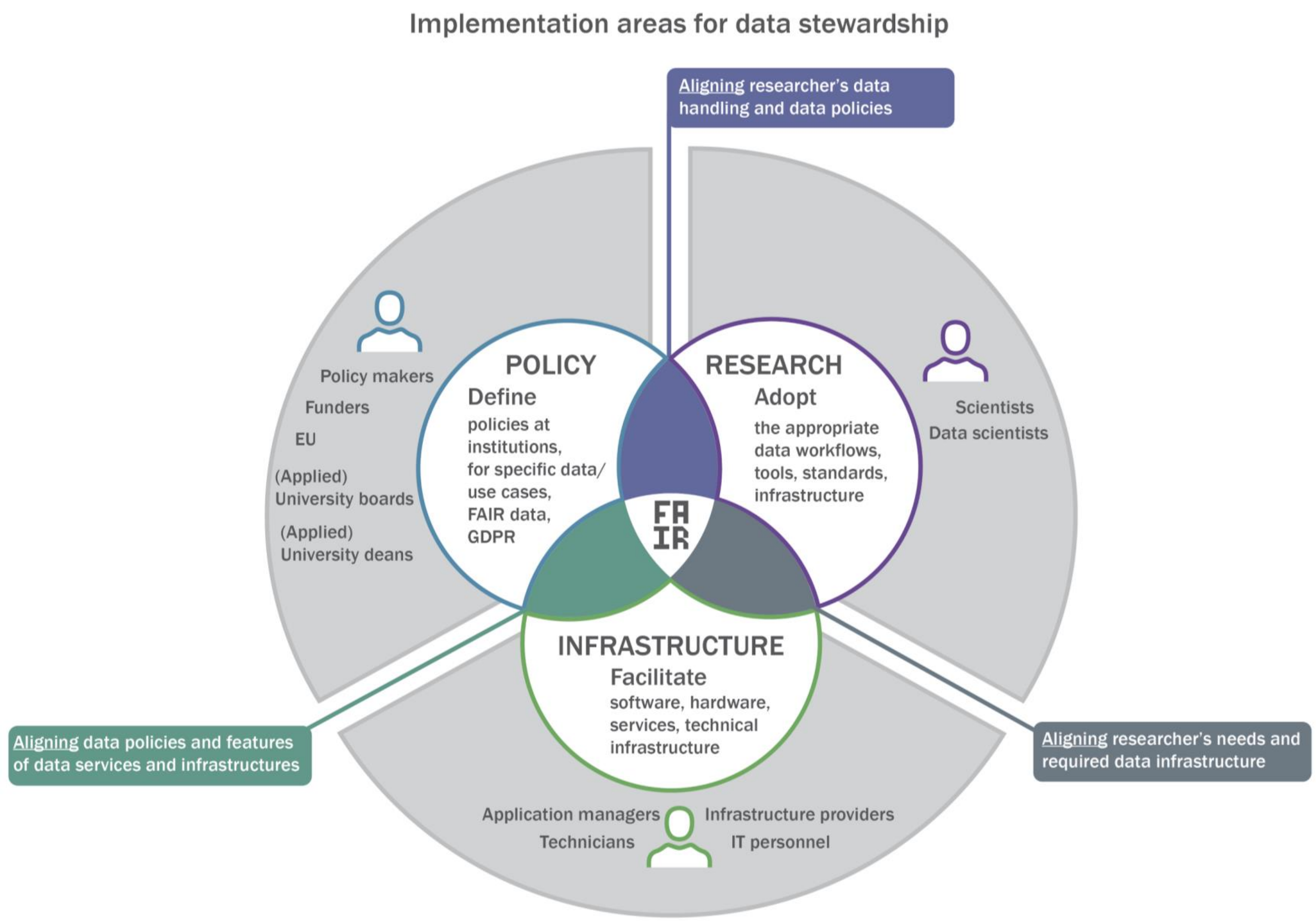
Research Data Lifecycle

Love My Data week
28th of March 2025

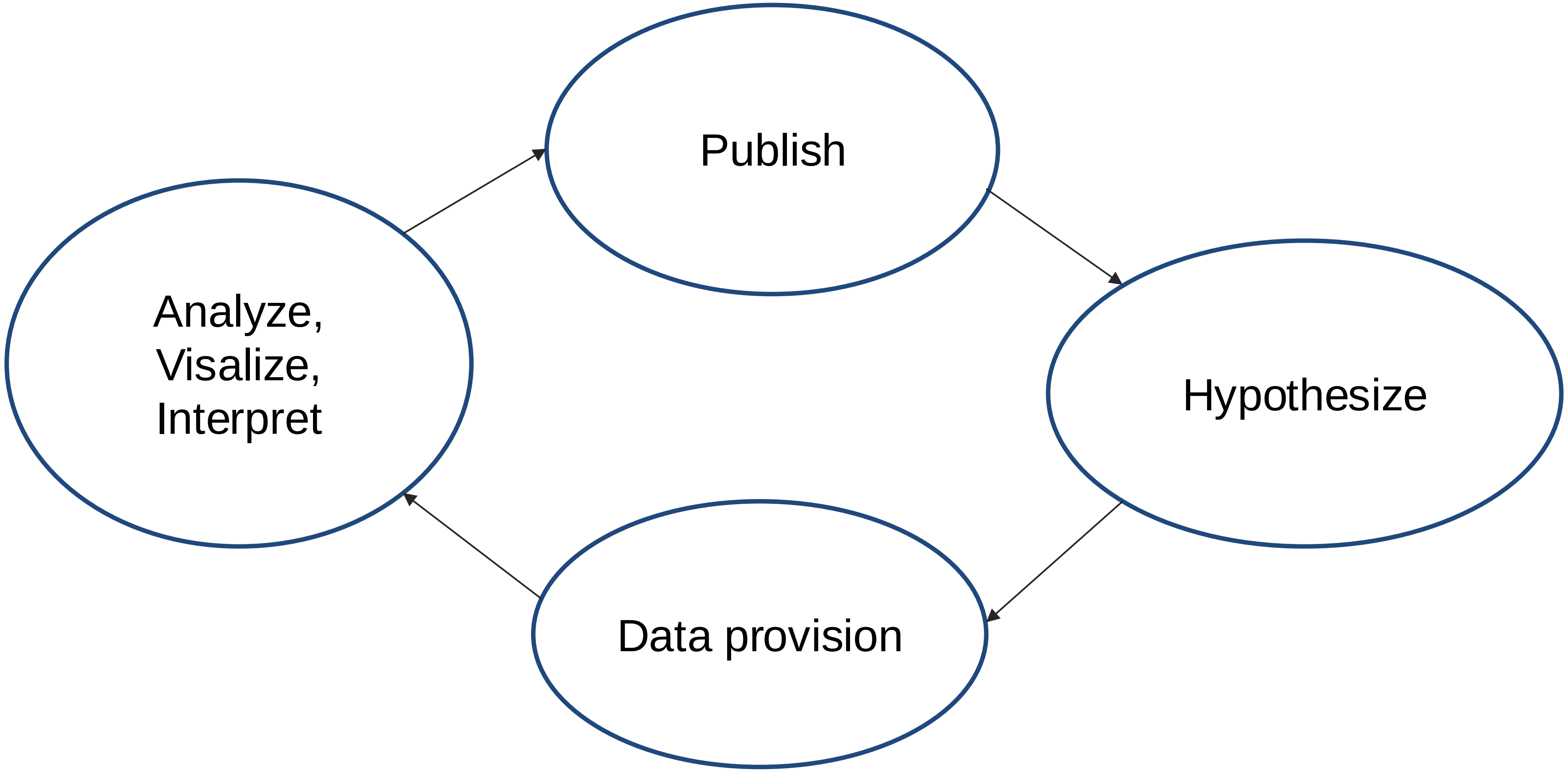
Vilem Ded

Data stewardship

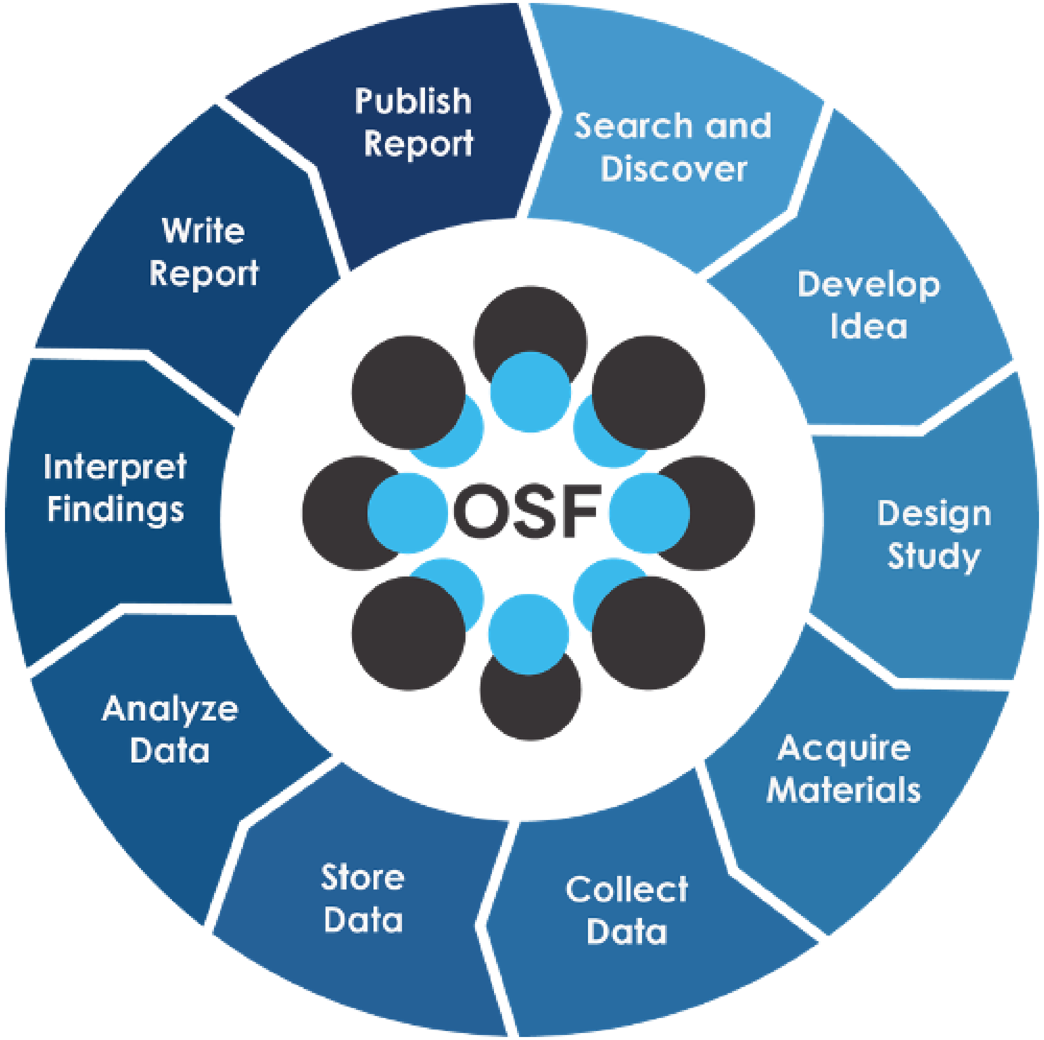

Implementation areas for data stewardship
Aligning researcher's data handling and data policies
Policy makers
Funders
EU
(Applied) University boards
(Applied) University deans
POLICY
Define
policies at institutions, for specific data/ use cases, FAIR data, GDPR
RESEARCH
Adopt
the appropriate data workflows, tools, standards, infrastructure
Scientists
Data scientists
FAIR
INFRASTRUCTURE
Facilitate
software, hardware, services, technical infrastructure
Aligning data policies and features of data services and infrastructures
Aligning researcher's needs and required data infrastructure
Application managers
Technicians
Infrastructure providers
IT personnel

Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M., & Van Gelder, C. W. G. (2021). Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship (1.1). Zenodo. https://doi.org/10.5281/zenodo.4623713

Research lifecycle

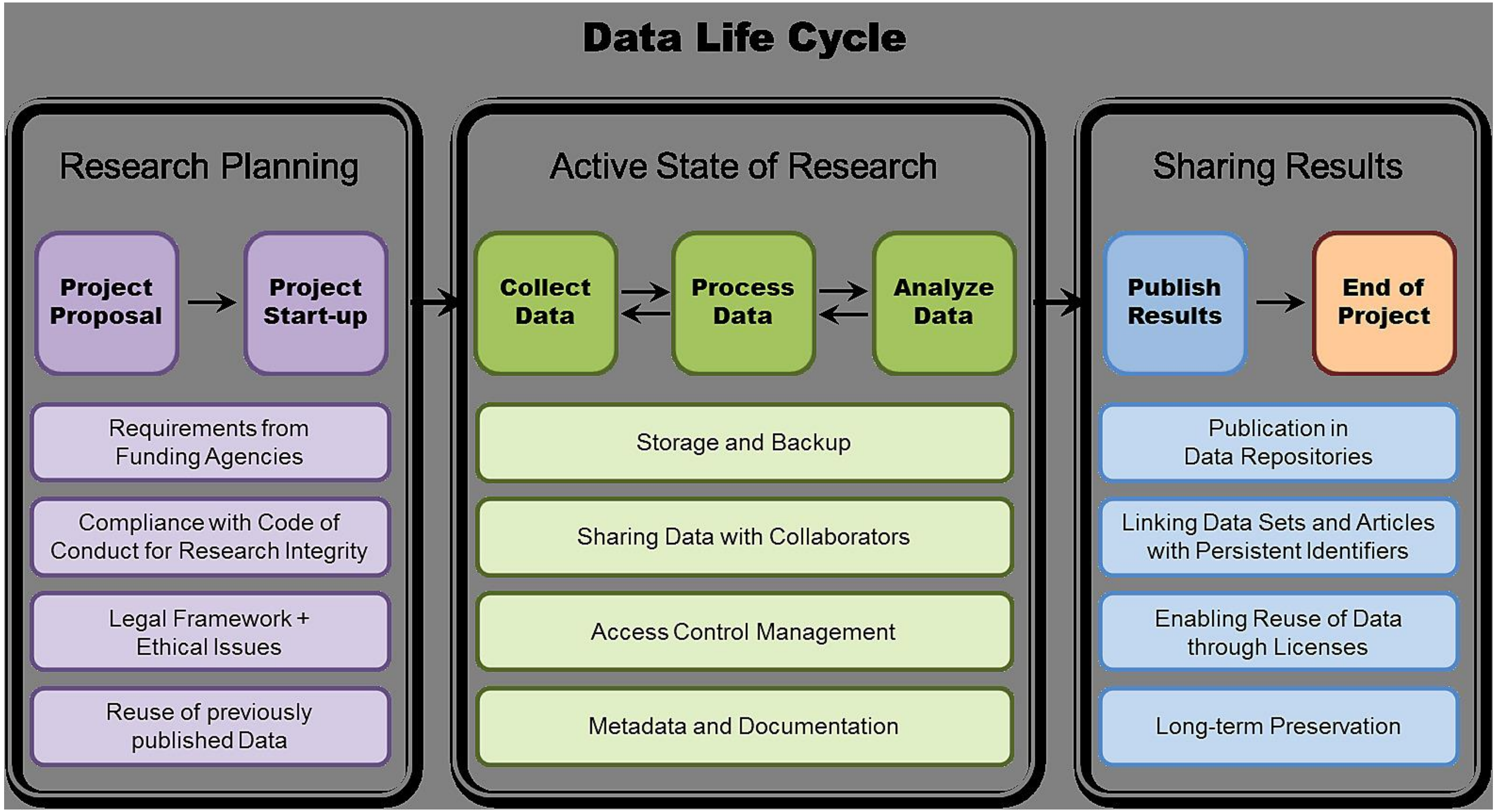


Research project lifecycle



Graphic provided by the Center for Open Science

Data lifecycle - 2016



Data Life Cycle

Research Planning

Project Proposal → Project Start-up

- Requirements from Funding Agencies
- Compliance with Code of Conduct for Research Integrity
- Legal Framework + Ethical Issues
- Reuse of previously published Data

Active State of Research

Collect Data ⇄ Process Data ⇄ Analyze Data

- Storage and Backup
- Sharing Data with Collaborators
- Access Control Management
- Metadata and Documentation

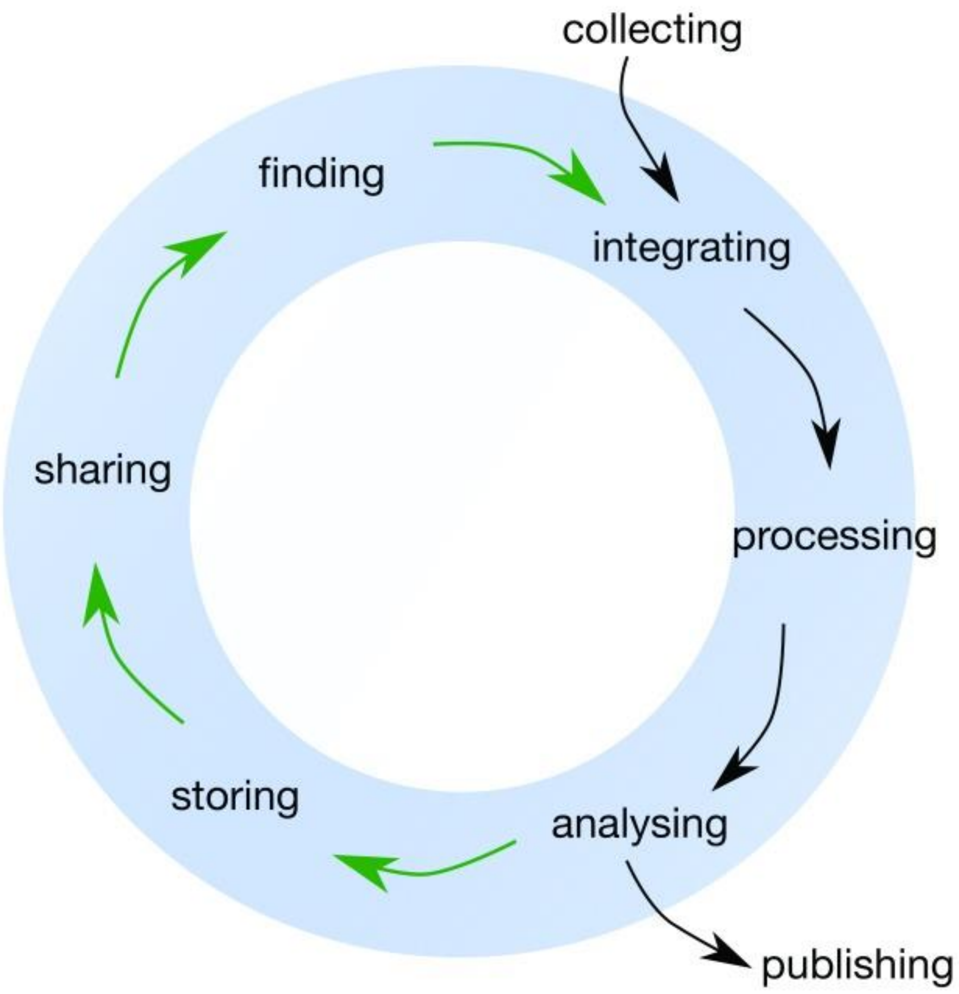
Sharing Results

Publish Results → End of Project

- Publication in Data Repositories
- Linking Data Sets and Articles with Persistent Identifiers
- Enabling Reuse of Data through Licenses
- Long-term Preservation

Hüser, Falco Jonas; Elbæk, Mikael K.; Martinez lavanchy, Paula (2016). DTU Research Data Life Cycle. Technical University of Denmark. Figure. https://doi.org/10.6084/m9.figshare.4258019.v1

Making lifecycle a cycle - 2018



Griffin PC, Khadake J, LeMay KS, Lewis SE, Orchard S, Pask A, Pope B, Roessner U, Russell K, Seemann T, Treloar A, Tyagi S, Christiansen JH, Dayalan S, Gladman S, Hangartner SB, Hayden HL, Ho WWH, Keeble-Gagnère G, Korhonen PK, Neish P, Prestes PR, Richardson MF, Watson-Haigh NS, Wyres KL, Young ND, Schneider MV. Best practice data life cycle approaches for the life sciences. F1000Res. 2017 Aug 31;6:1618. doi: 10.12688/f1000research.12344.2. PMID: 30109017; PMCID: PMC6069748.

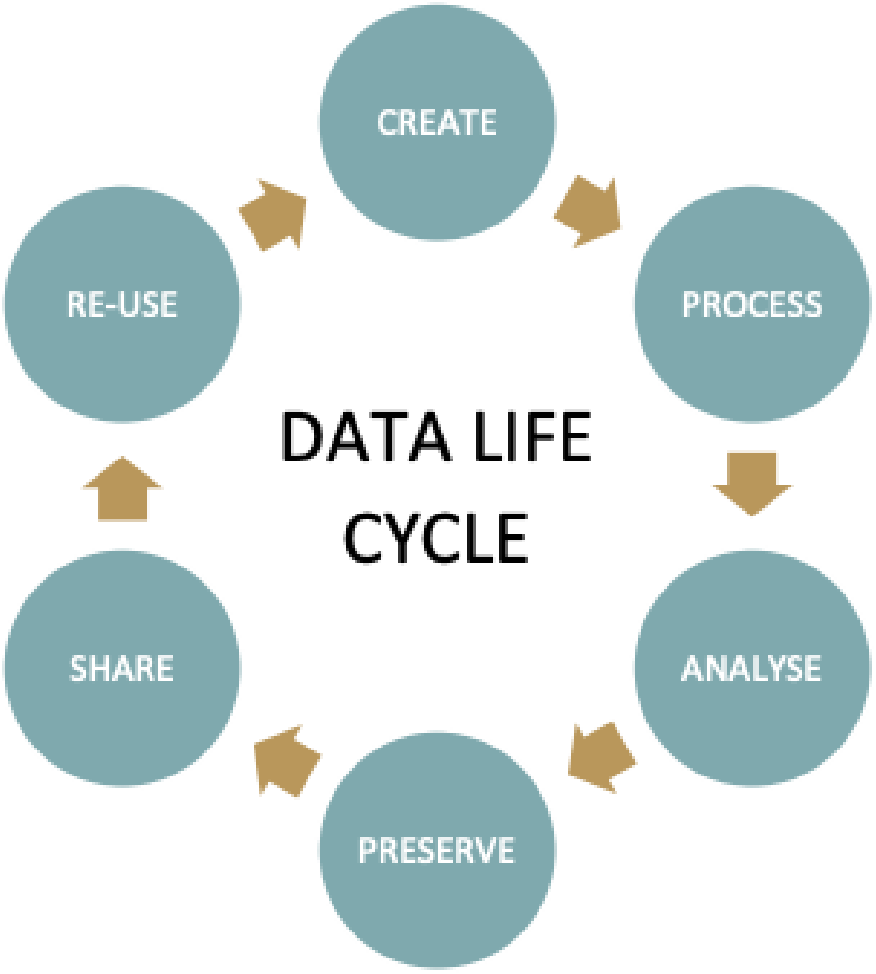
FNR and research data management

Luxembourg National Research Fund

FNR POLICY ON RESEARCH DATA MANAGEMENT

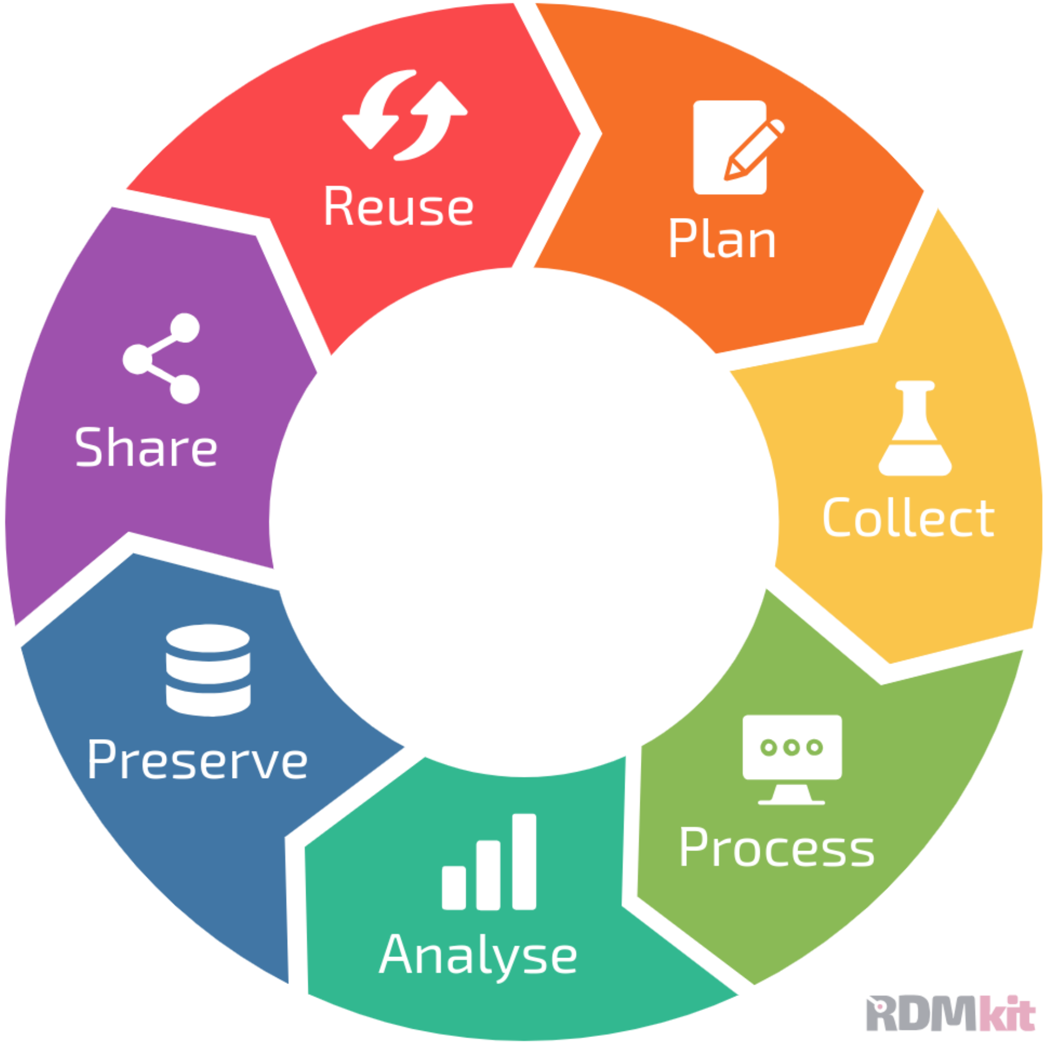
Research Data Management *(RDM) is a general term covering how you organise, structure, store, and care for the data used or generated during the lifetime and after the completion of a research project. It is good research practice to ensure that your data are managed properly throughout the life of the project. This means planning how you will collect, store, and care for your data before you start the research process, through to how you will ensure it is maintained in the longer-term and shared with your research community and beyond.*

Further benefits can be derived from good data management, including accessibility, sustainability, impact, speed, efficiency.



Research Data Lifecycle is a concept which *provides a broader view of the stages data goes through (during a research project)*

Research data lifecycle



"[Data life cycle diagram]" by [RDMkit] / [CC BY]

ELIXIR (2021) Research Data Management Kit (RDMkit). A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075). URL: https://rdmkit.elixir-europe.org

Plan

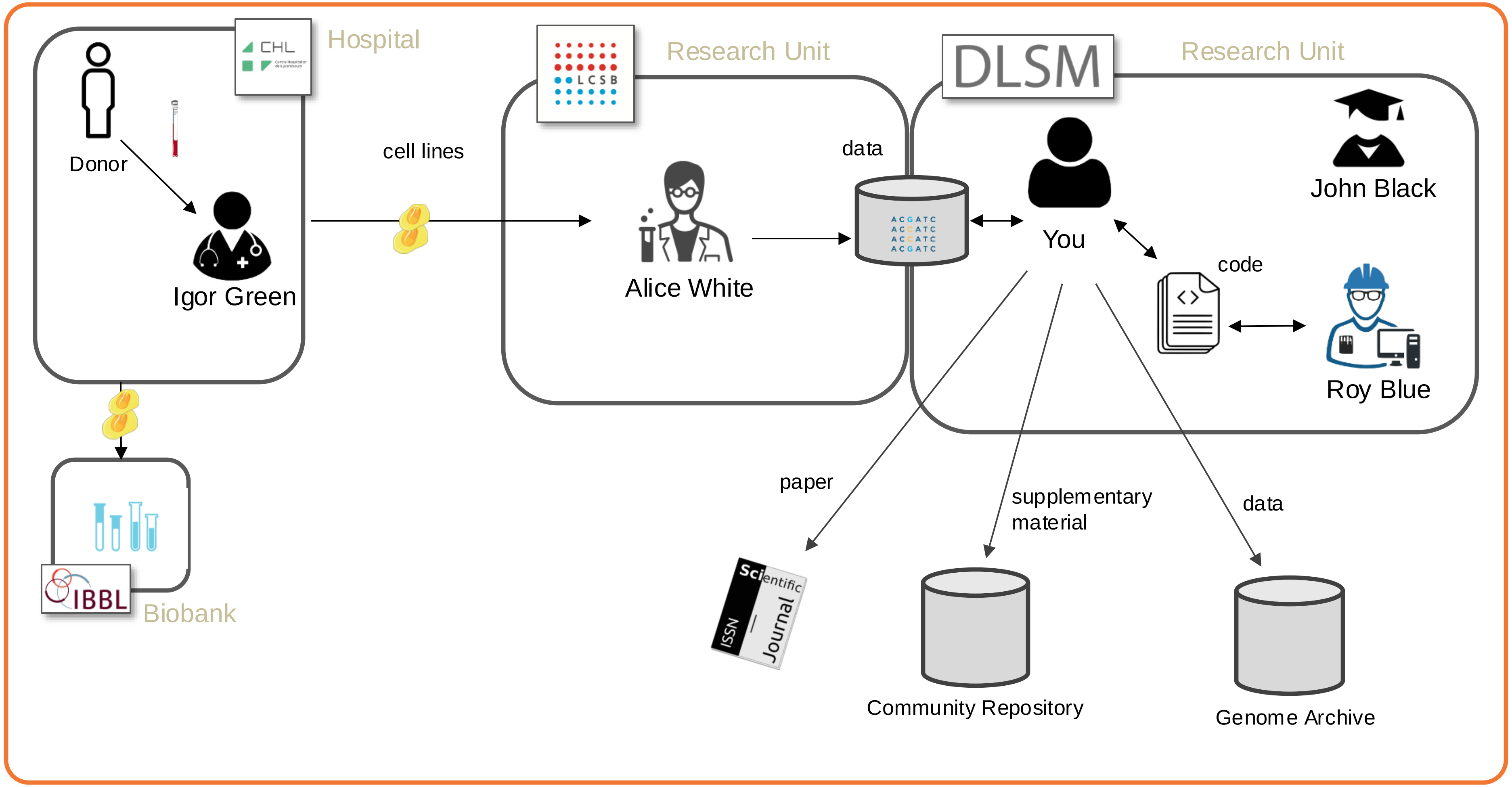
Data Management Plan (DMP)

- A formal document that outlines what to do regarding data before, during and after a research project
- Requirement of research organisations and funders
- Living document: researchers are accountable for how data is treated. You consult it during work and **change** it when needed (review, new data, change in policy, change in consortium agreement, …)

UNI instance of DMPonline
https://unilu.dmponline-mt.dcc.ac.uk/
FNR template
Simple web based form

ELIXIR-LU DS Wizard
https://elixir-lu.ds-wizard.org/
FNR template
More detailed questionnaire
DPIA model and project template

Workflow diagram



Collect

Collecting data

Collecting existing data - requires a legal framework (license, contract, participation in a project, ...)

Generate data – experiments, interviews, synthetic datasets, recordings, ...

Quality control and assurance

- Dropdown list
- Avoid open text fields
- Cross validate with data dictionary
- Create detailed data collection protocol
- Ask for peer review
- ...

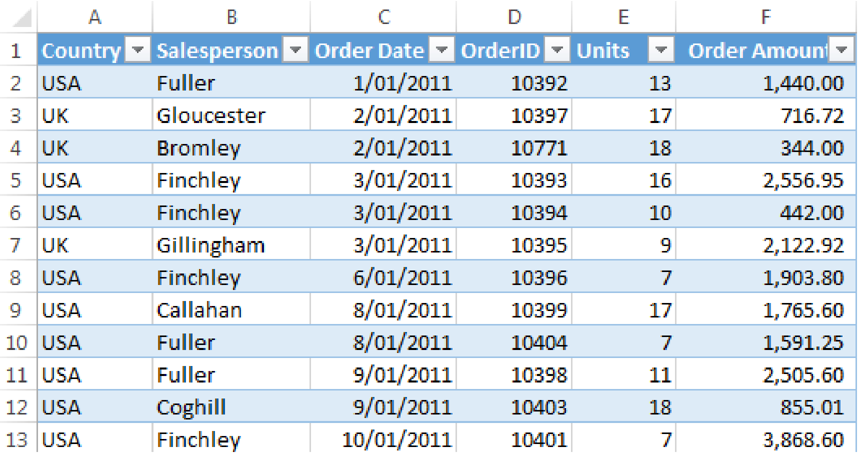
Ingestion procedure

- Validate the data is what it is supposed to be
- Check integrity
- Move it to data warehouse
- Revoke access
- Create a record in data catalog
- ...

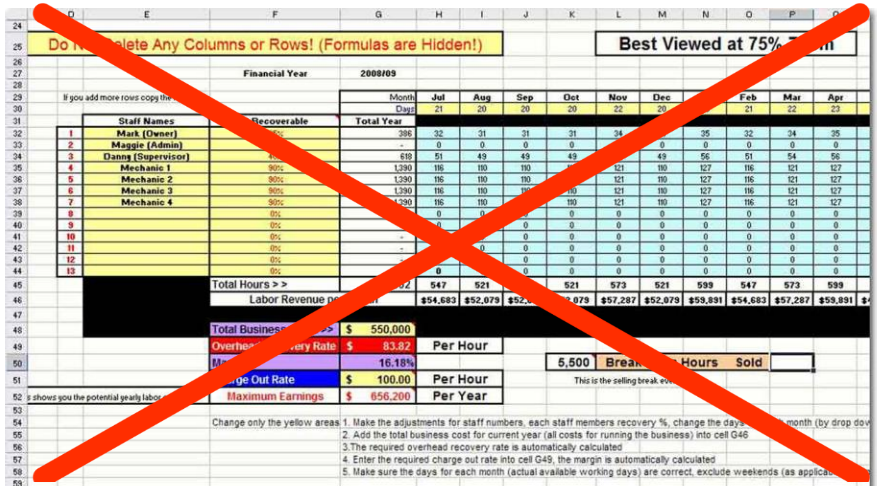
Working with spreadsheets

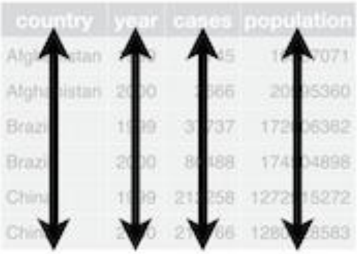
Tabular format:

1. **One-line header**
   - Unique and machine-readable column names
2. **Rows**
   - Represent individual observations/entities
3. **Columns**
   - Represent attributes/features of the observations
   - Contain values of one data type

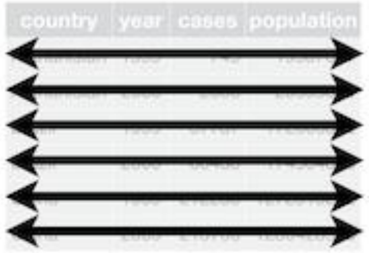


	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60

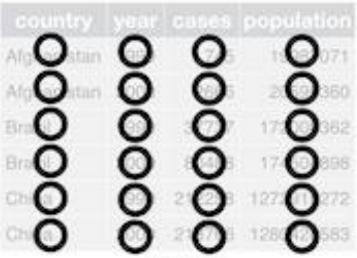




variables



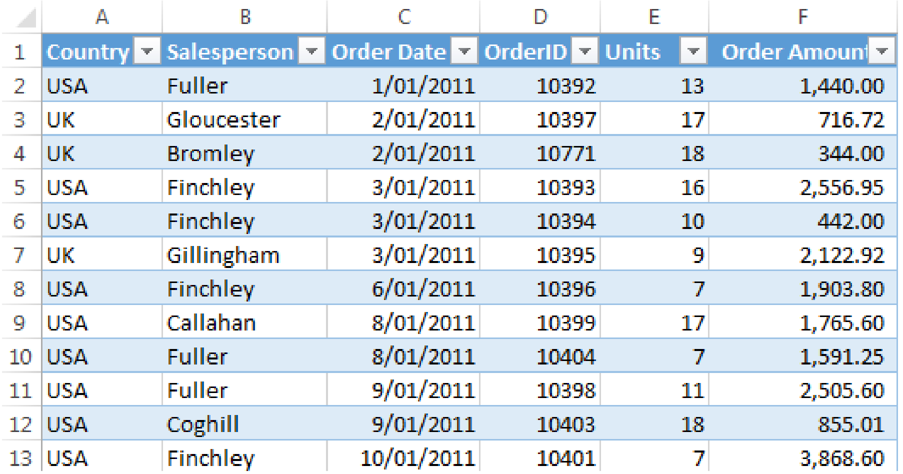
observations



values

Data dictionary

Data



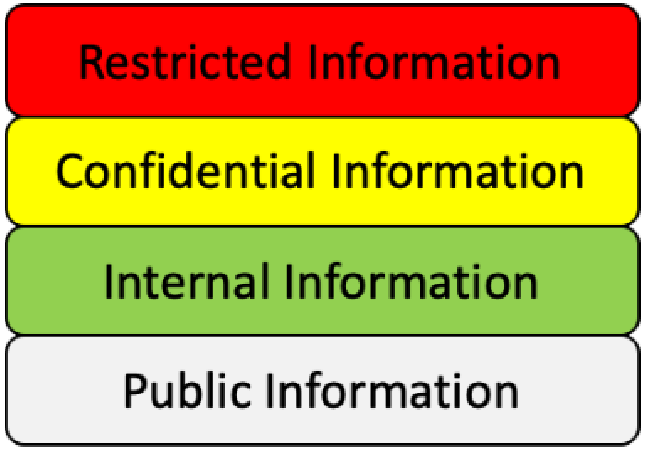
	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60

Data dictionary

column	data_type	description	value_min	value_max
Country	text	Company branch receiving the order	N/A	N/A
Salesperson	text	Surname of the person responsible for the sale	N/A	N/A
Order Date	date	Date on which the order was submitted to the system	1900-01-01	2020-01-01
OrderID	number	Unique identifier of the order (see ORDERS table for more details)	0	100 000
Units	number	the number of individual items that a company sells	0	100
Order Amount	number	All purchase prices referenced in any purchase order(s) and thus the total, accumulated and aggregated sum of each and every purchase price	1	10 000

Data classification and cataloging

Information security

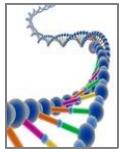

Restricted Information

Confidential Information

Internal Information

Public Information

GDPR

	Peter Pan	Patient X
	Identified Sensitive	Pseudonymised Sensitive
	Identified	Pseudonymised

Process

Data (pre-)processing

data analysis

data pre-processing

Hypothesis testing
Visualization
Training AI model

Cleaning
Validating
Harmonizing
Summarizing
Merging
Unifying
Splitting
Transforming
Standardizing
Parsing
Enriching

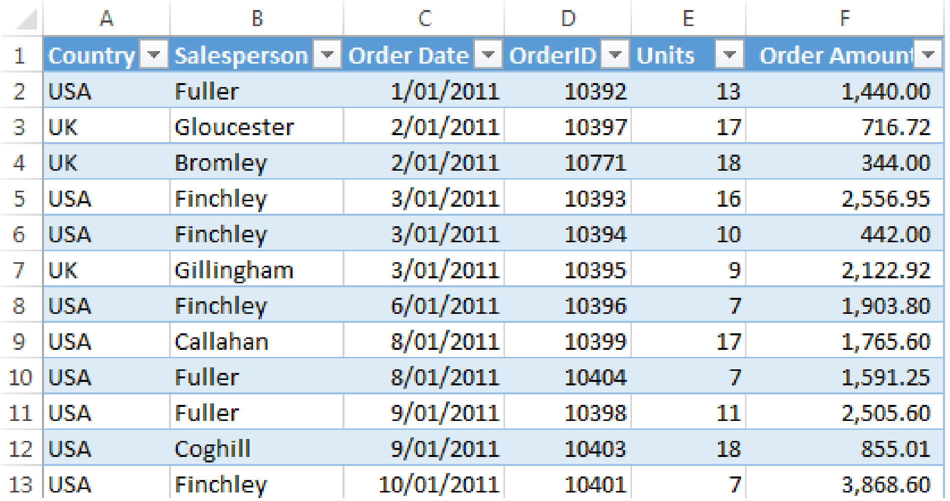
Reproducibility

Spreadsheets alone

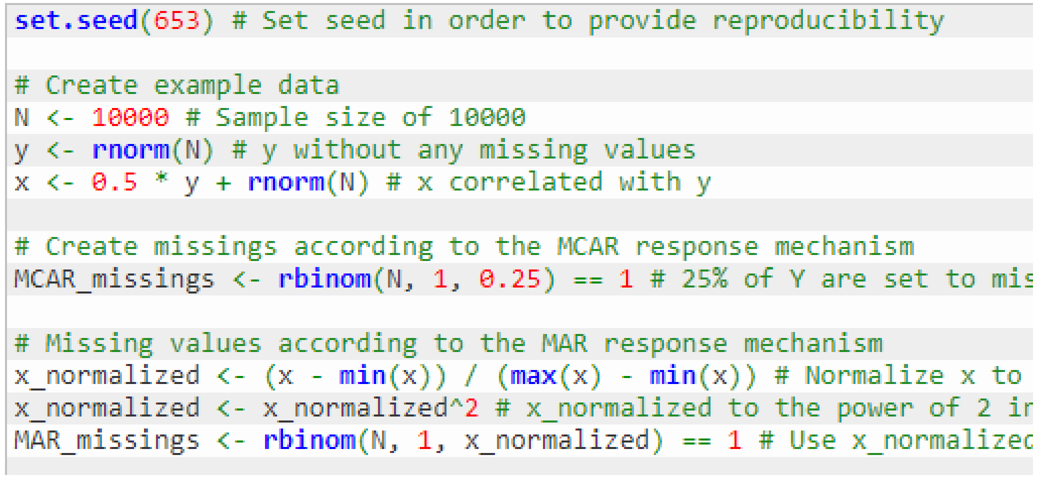
- Is great for looking at data.
- Data entry is fast.
- Analysis flow is hidden and not in focus.

Coding

- Is great for controlling analysis
- Data is hidden.
- Flow is visible.



	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60



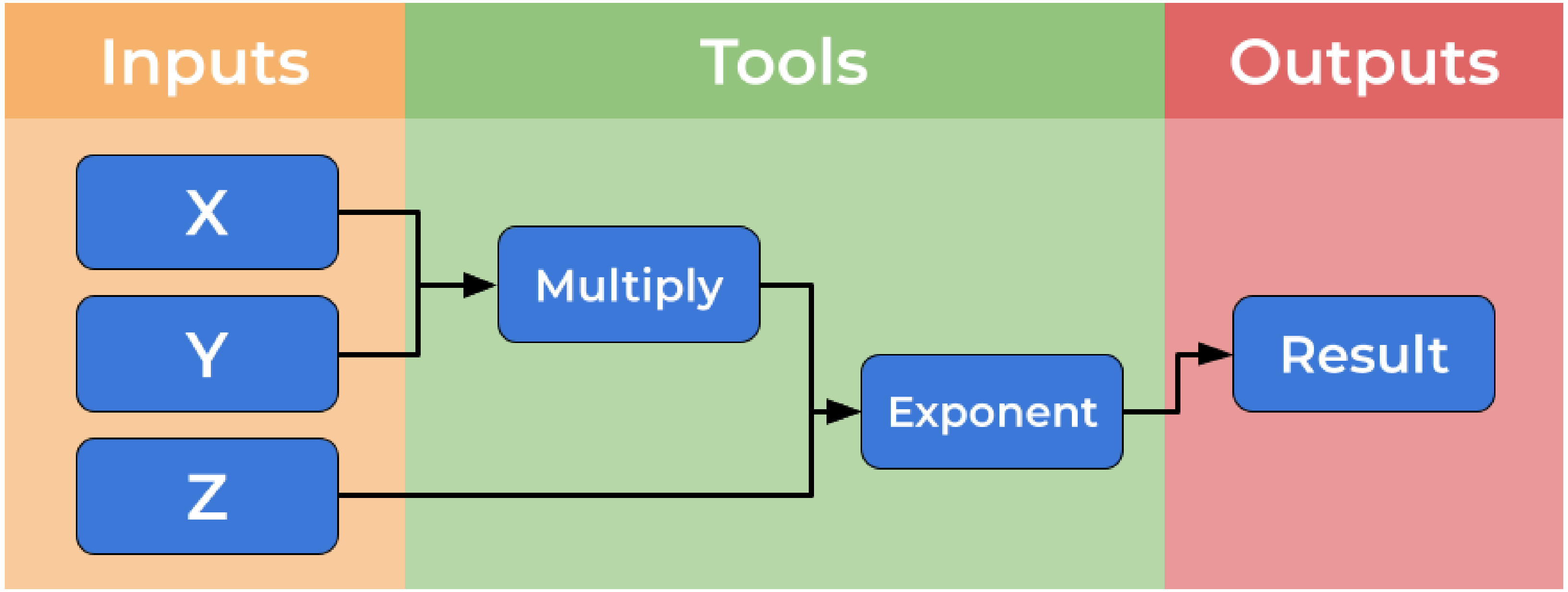
```r
set.seed(653) # Set seed in order to provide reproducibility

# Create example data
N <- 10000 # Sample size of 10000
y <- rnorm(N) # y without any missing values
x <- 0.5 * y + rnorm(N) # x correlated with y

# Create missings according to the MCAR response mechanism
MCAR_missings <- rbinom(N, 1, 0.25) == 1 # 25% of Y are set to mis

# Missing values according to the MAR response mechanism
x_normalized <- (x - min(x)) / (max(x) - min(x)) # Normalize x to
x_normalized <- x_normalized^2 # x_normalized to the power of 2 ir
MAR_missings <- rbinom(N, 1, x_normalized) == 1 # Use x_normalizec
```

Capturing changes, computational workflows and dependencies



[Carpentry course on Introduction to Workflows with Common Workflow Language](https://carpentries-incubator.github.io/cwl-novice-tutorial/aio/index.html)
https://carpentries-incubator.github.io/cwl-novice-tutorial/aio/index.html

renv

COMMON WORKFLOW LANGUAGE

docker

CONDA®

snakemake

Galaxy

GitLab GitHub

Analyze

Preserve

Storage != Preservation

Preservation

"The act of conserving and maintaining both the safety and integrity of data."Wikipedia

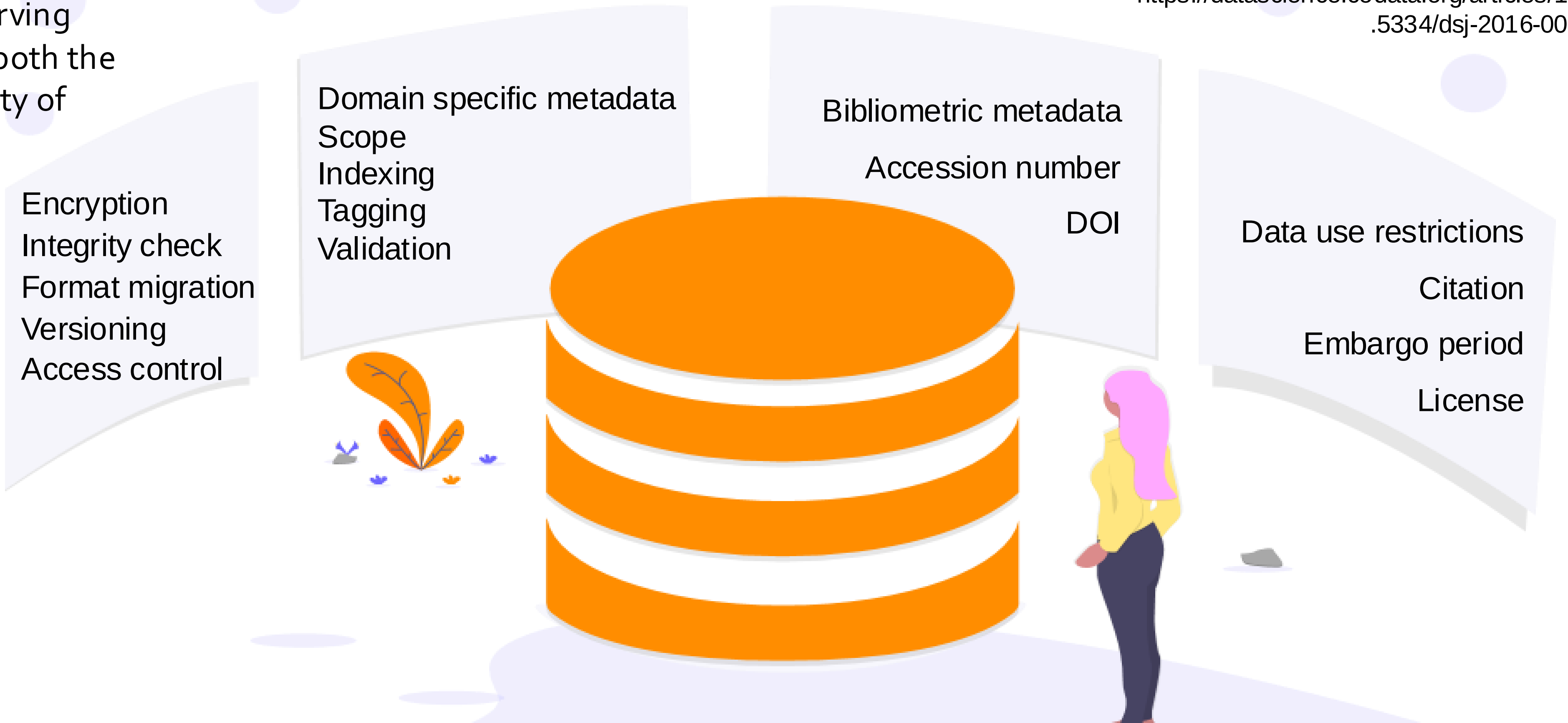
Archival

The process of moving data that is no longer actively used to a separate storage device for long-term retention.

Publication

"Research data publishing is an approach for **sharing research data**, i.e., it is intended as the release of (research) data **for (re)use** by others. "

https://datascience.codata.org/articles/10.5334/dsj-2016-006



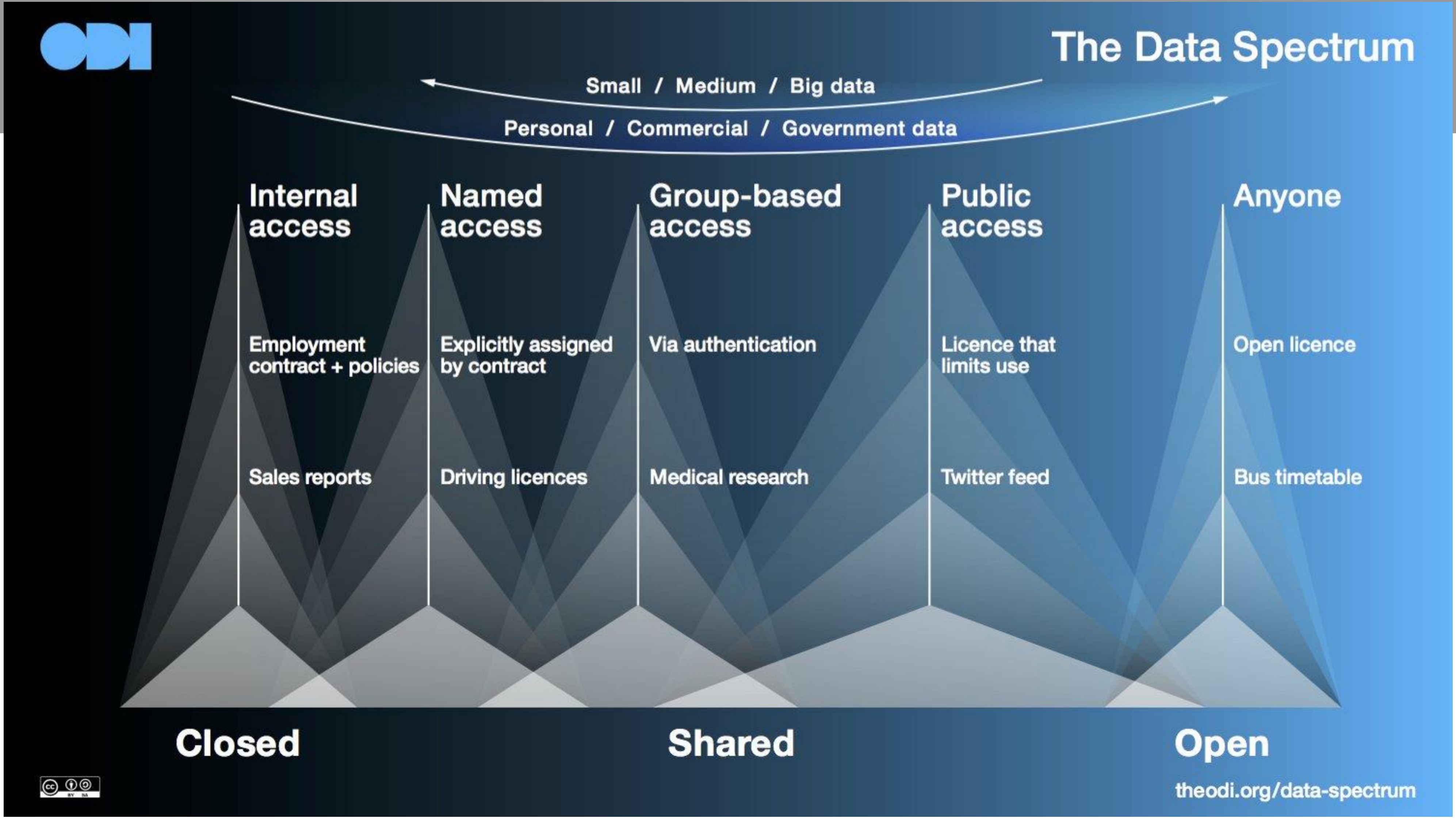
Encryption
Integrity check
Format migration
Versioning
Access control

Domain specific metadata
Scope
Indexing
Tagging
Validation

Bibliometric metadata
Accession number
DOI

Data use restrictions
Citation
Embargo period
License

Share



The Data Spectrum

Small / Medium / Big data

Personal / Commercial / Government data

Internal access	Named access	Group-based access	Public access	Anyone
Employment contract + policies	Explicitly assigned by contract	Via authentication	Licence that limits use	Open licence
Sales reports	Driving licences	Medical research	Twitter feed	Bus timetable

Closed Shared Open

theodi.org/data-spectrum

You start here → You want/have to get here

Challenges when sharing data

Phil Archer Diary 2015: https://philarcher.org/diary/2015/50shadesofno/

Too expensive

There’s no business case

There’s no commercial value

It’s private

It’s secret

It's our data

We have invested a lot of money in this

Link enough data and one will arrive at sensitive private information

It's not data, it's information

It will never work

We don't know how to do this

We don't have the right people to do this

We need the money

It’s not ours, and we don’t know who’s data it is

No idea what the quality of the data is

We don’t know where to find it

It’s not our job

It isn’t in the right format

I am not authorised

Who is going to use this anyway

People are going to misuse it

Image damage for the minister

We are not ready for this

Image loss for Government

The data file is too big

Not enough bandwidth

This is a first step, we will see what we can do later

We can’t find it

We have no access

It is out of date / too old

We have it only on paper

We don’t know if it’s legal

Management says no

We never did this before

No value in it

No time / no resources

We will open up (but adapt 90%)

It’s incorrect

Commercially sensitive

It is dangerous when linked

People are going to make the wrong conclusions

This is going to start a wrong discussion

We can’t say whether we have it or we don’t

We know the data is wrong, and people will tell us where it is wrong, then we'd waste resources inputting the corrections people send us

Our IT suppliers will charge us a fortune to do an ad hoc data extract

We have to be careful withs existing contracts

Our website cannot hold files this large

It's not ours and we don't have authorization from the data owner

We've already published the data (but it's unfindable/unusable)

People may download and cache the data and it will be out of date when they reuse it

We don't collect it regularly

Too many people will want to download it, which will cause our servers to fail

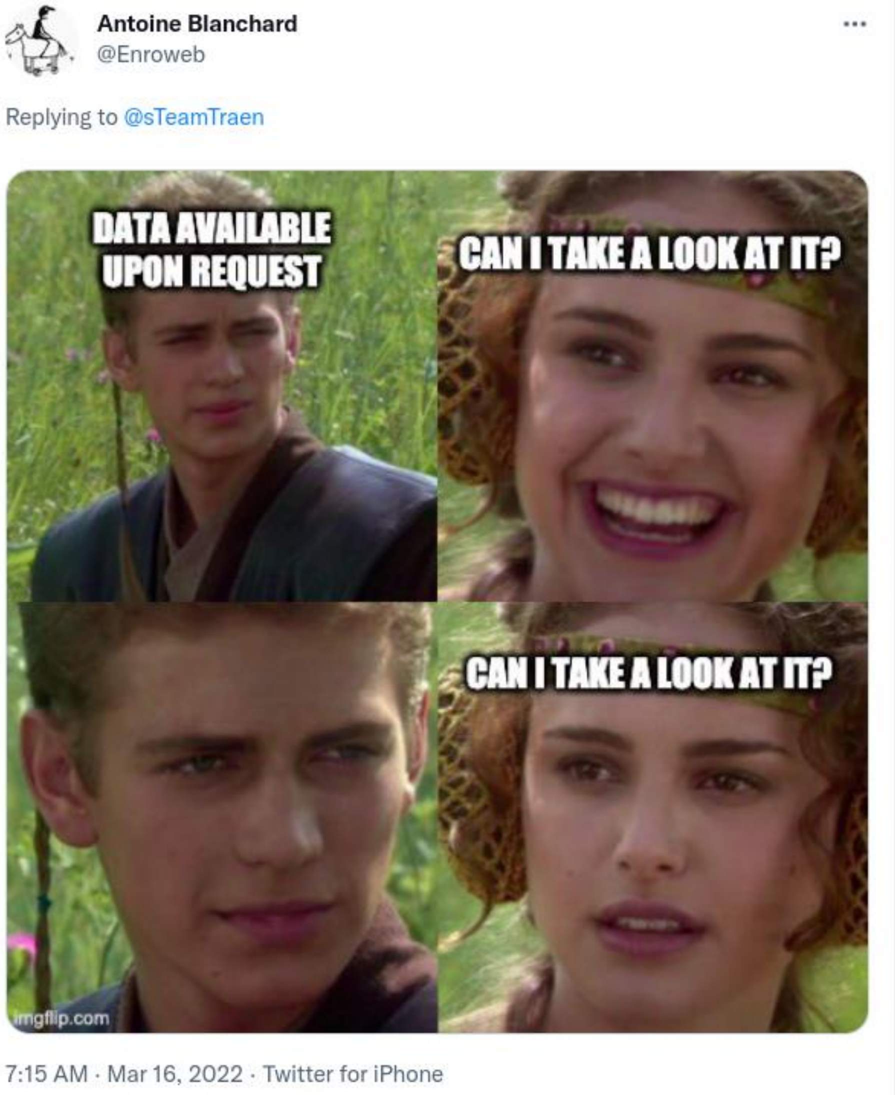
People would get upset

It’s very sensitive information

We are not ready for this

Tell us who is going to use it and we will make it open

Data availability statement



Antoine Blanchard
@Enroweb

Replying to @sTeamTraen

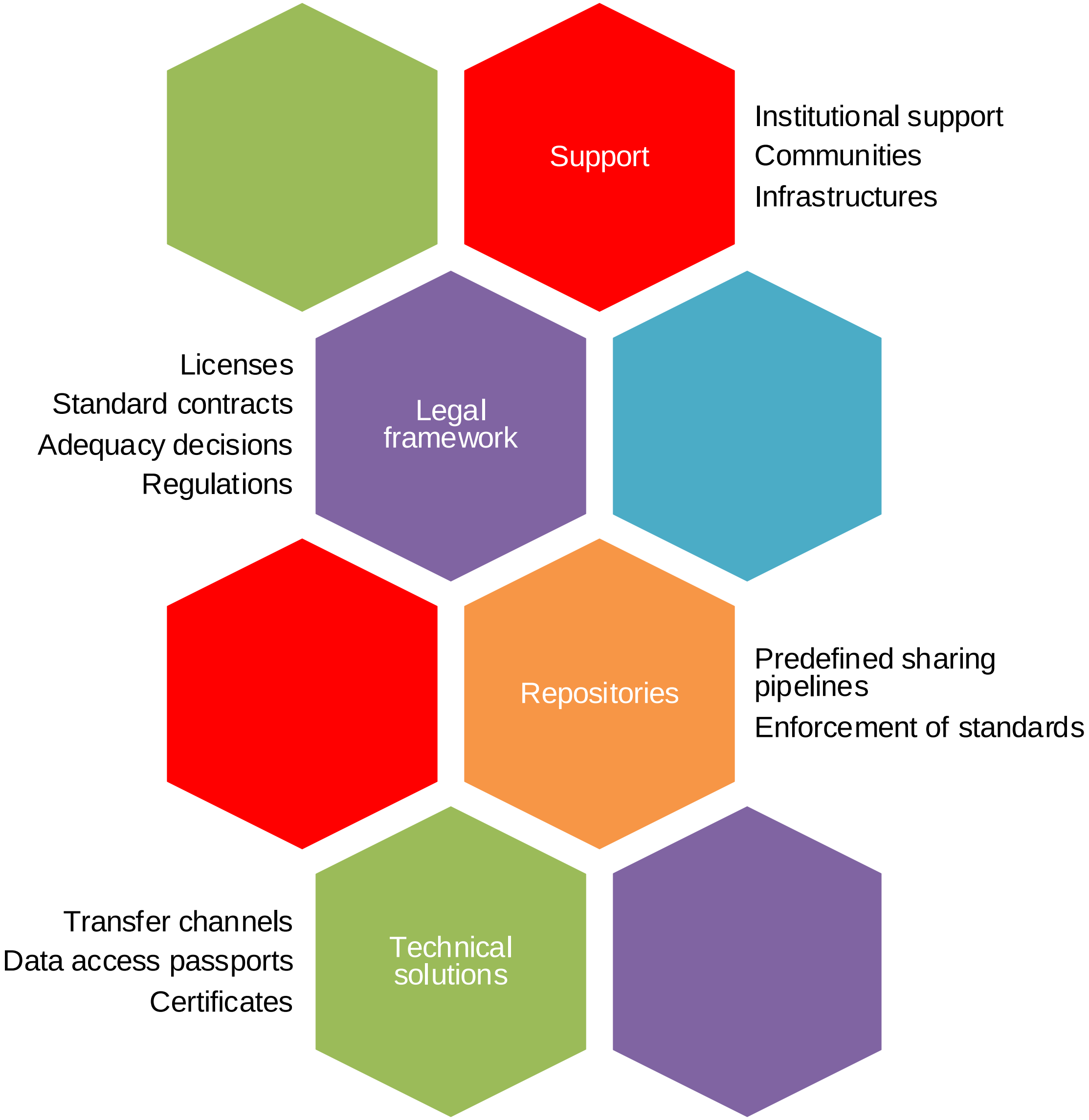
7:15 AM · Mar 16, 2022 · Twitter for iPhone

Antoine Blanchard on Twitter: "@sTeamTraen https://t.co/MpQj6MYK8r

FNR Open Access Policy:

"In no way is it acceptable to merely include a simple statement "data available on request" or similar."

Data sharing made ~~easy~~ easier



Support

Institutional support
Communities
Infrastructures

Legal framework

Licenses
Standard contracts
Adequacy decisions
Regulations

Repositories

Predefined sharing pipelines
Enforcement of standards

Technical solutions

Transfer channels
Data access passports
Certificates

Re-use

Re-use data!

One of the FAIR principles

Benefits:

- obtain reference data for your research
- **avoid** doing new, **unnecessary experiments**;
- run analyses to verify that reported findings are correct, and thereby making subsequent **findings more robus**t;
- make **research more robust** by aggregating results obtained from different methods or samples;
- gain novel insights by **connecting** and meta-analysing datasets.

How to find data of interest?

- Ask around ☺
- Search for/in specific repository
- Use specialized search tools:

  Elsevier Data Search

  Google Dataset

  Data Citation Index

  Data Discover Index

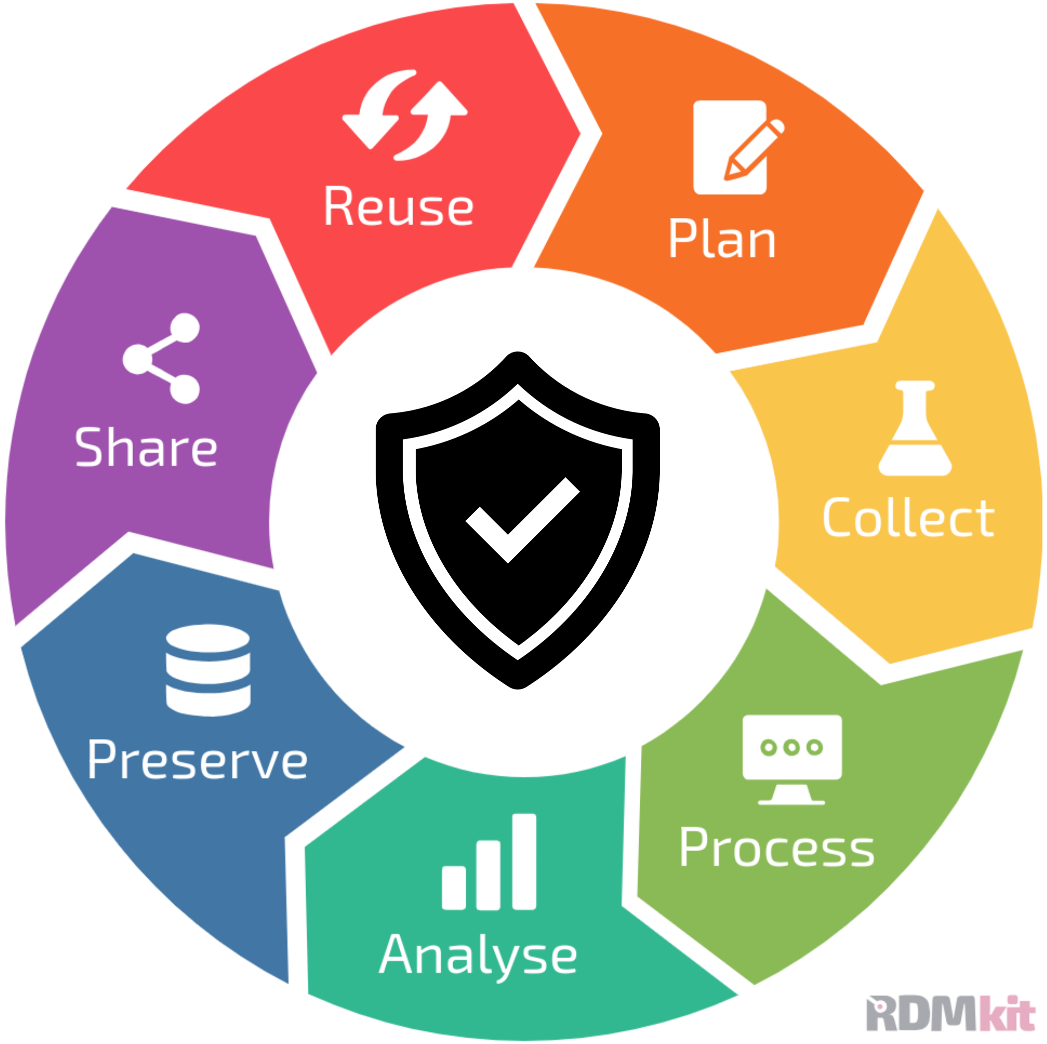
There is more ☺

Data lifecycle – missing dimensions

Data Security

This primarily involves safeguarding data from unauthorized access, breaches, and cyber threats.

- Encryption
- Firewalls
- Access controls
- ...



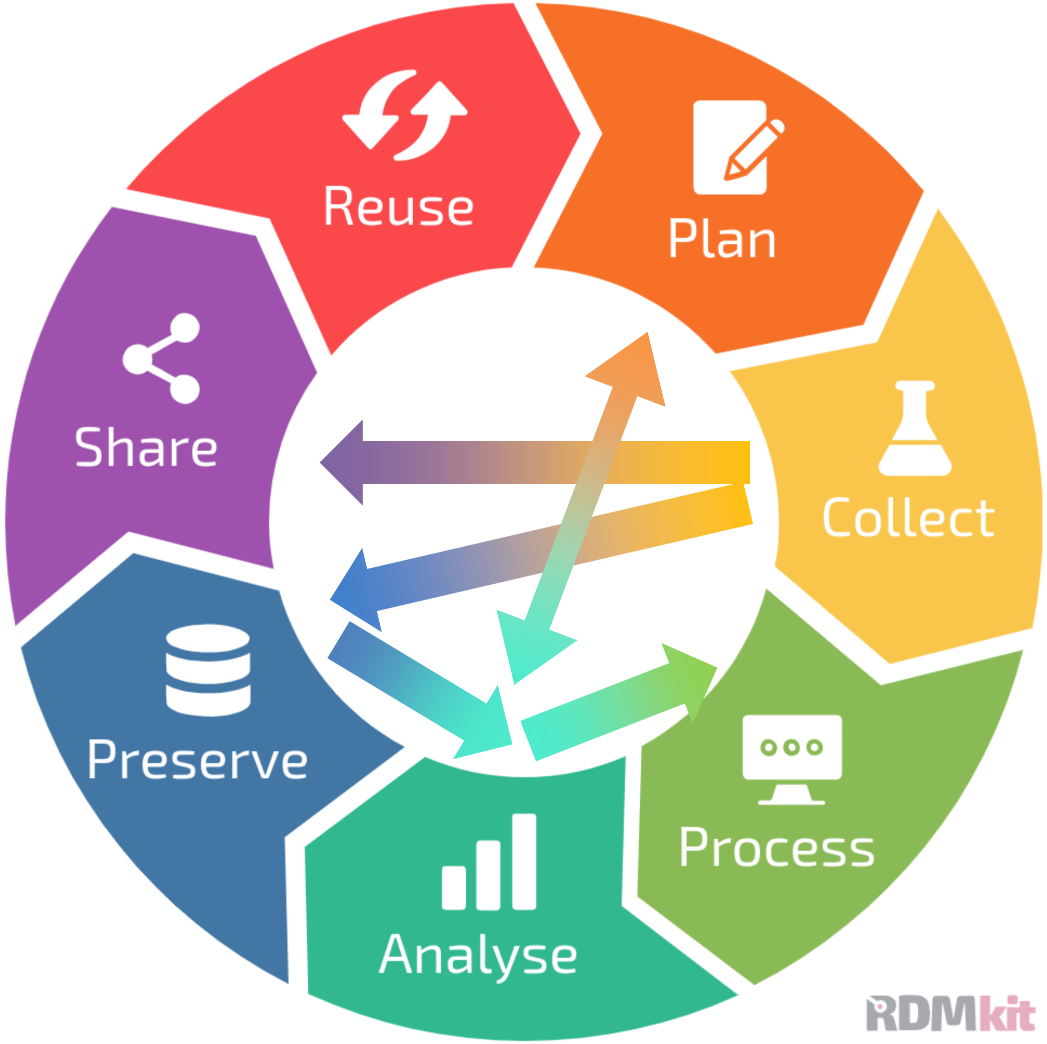
Data Protection

This has a broader scope, encompassing not only the security of data but also its lawful and ethical use.

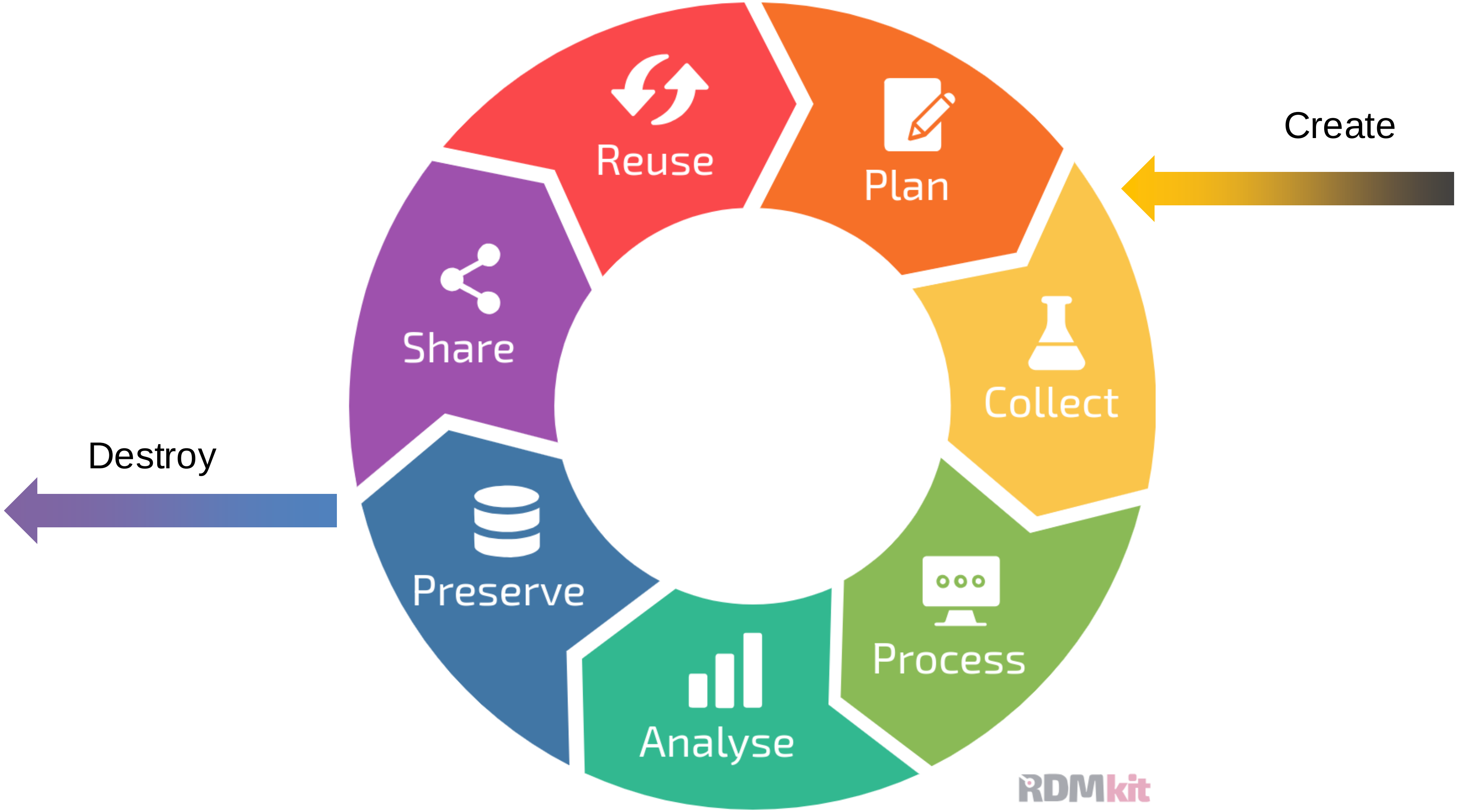
- Ethics
- GDPR
- Contractual frameworks
- ...

"Data life cycle diagram" by RDMkit / CC BY
ELIXIR (2021) Research Data Management Kit (RDMkit). A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075). URL: https://rdmkit.elixir-europe.org

Data lifecycle - unidirectionality



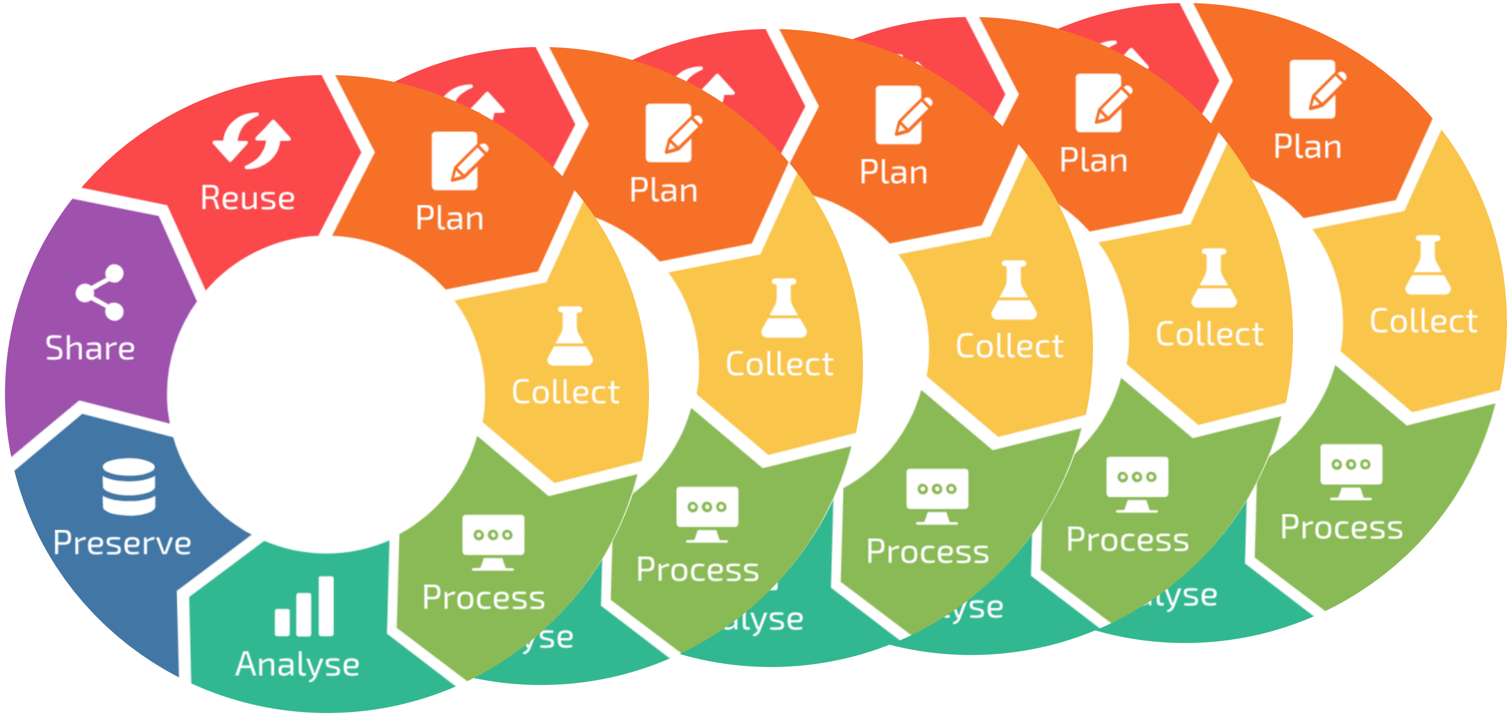
Data lifecycle - Closeness



Scope of the lifecycle

UNIVERSITÉ DU LUXEMBOURG

Every data asset has its own life



Does it include metadata?

ELIXIR and ELIXIR Luxembourg

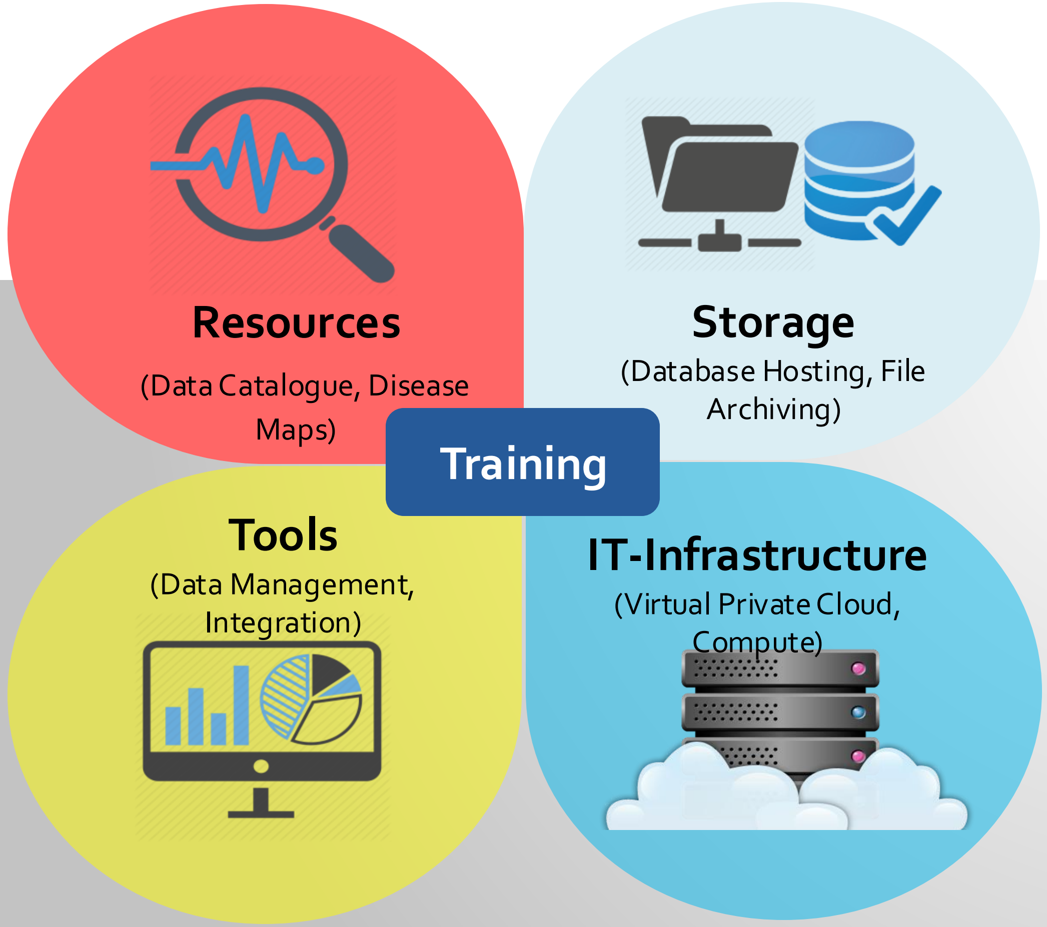
ELIXIR

European bioinformatics research infrastructure (S3)

National nodes

www.elixir-europe.org
www.elixir-luxembourg.org

ELIXIR Luxembourg services



Resources
(Data Catalogue, Disease Maps)

Storage
(Database Hosting, File Archiving)

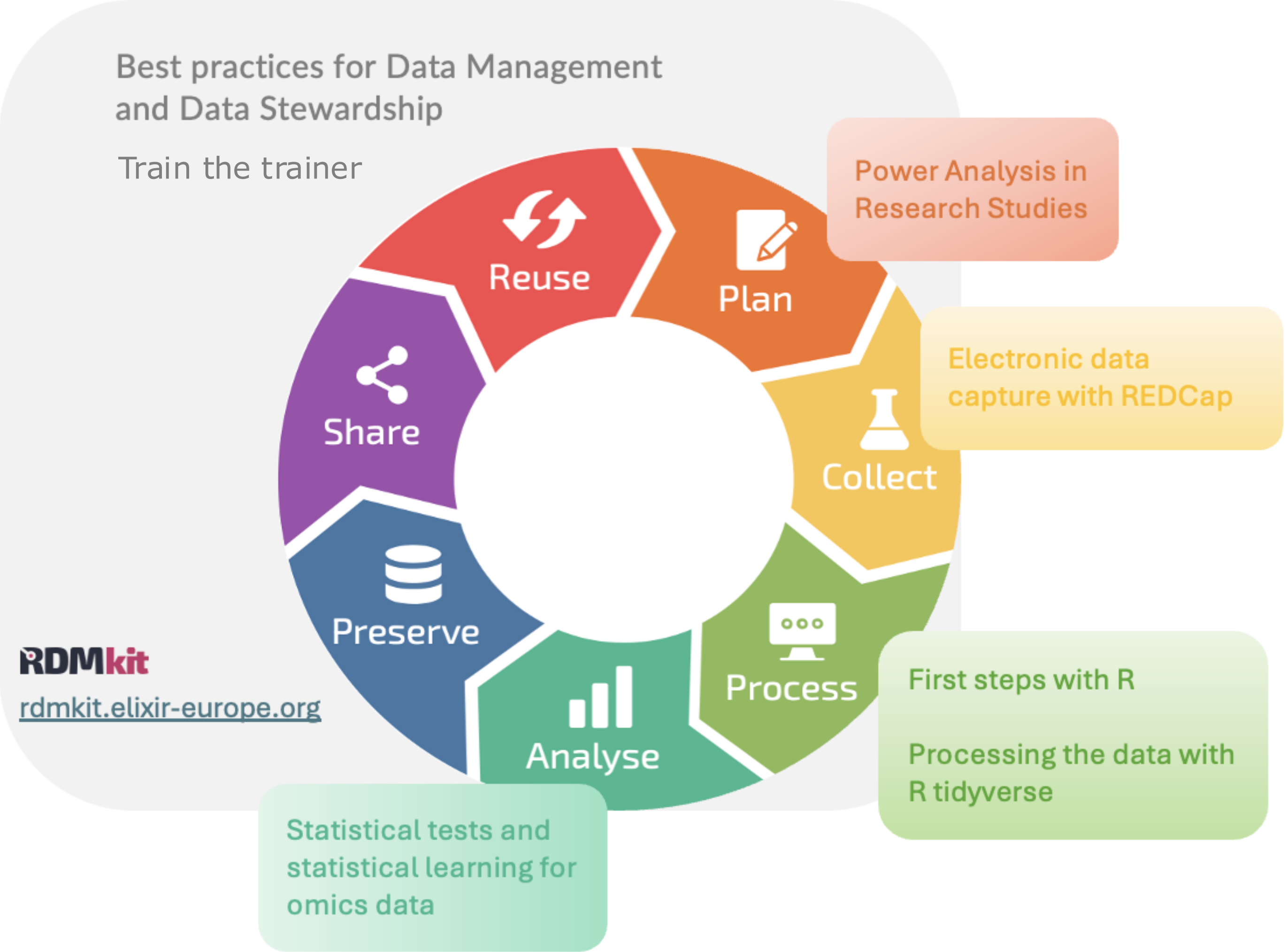
Training

Tools
(Data Management, Integration)

IT-Infrastructure
(Virtual Private Cloud, Compute)

Training in ELIXIR Luxembourg

- Support researchers with training concepts developed in ELIXIR
- Focus on Node’s mission
  - Data literacy
  - Data management and data stewardship
  - Scientific reproducibility
- Course management



Best practices for Data Management and Data Stewardship

Train the trainer

rdmkit.elixir-europe.org

Power Analysis in Research Studies

Electronic data capture with REDCap

First steps with R

Processing the data with R tidyverse

Statistical tests and statistical learning for omics data

www.elixir-luxembourg.org

External resources - starter kit

- RDMKit - https://rdmkit.elixir-europe.org/
- FAIR Cookbook - https://faircookbook.elixir-europe.org/
- The Turing way - https://the-turing-way.org
- …

Thank you!

UNIVERSITÉ DU
LUXEMBOURG