

Data Management

S2: Experimental data (1)-Storage

Monday 27/11/2023 - 14:00-16:00 (CET)

Isabelle Alic (INRAE)

Farzaneh Kazemipour-Ricci (INRAE)

Pascal Neveu (INRAE)

General objectives: Overview of data management for plant phenotyping - focus on FAIR data

Session 2

Experimental Data:

Storage & Data Management Policies

Overview

- Quick review of Session 1
- Data Storage - Introduction
- Cloud Computing
 - Data Storage Services
 - Data processing Services
- Data Management Policies
 - DMP - Data Management Plan
 - GDPR - General Data Protection Regulation

Overview

- Quick review of Session 1
- Data Storage - Introduction
- Cloud Computing
 - Data Storage Services
 - Data processing Services
- Data Management Policies
 - DMP - Data Management Plan
 - GDPR - General Data Protection Regulation

Phenotyping experiments : Dealing with a big mess!

- ❖ Complex & massive data: source, type, operator, scale, transformations, etc.
 - Breeding, genomic, weather, soil, observations, sensors, internet, etc
- ❖ Bad habits dealing with this mess
 - Dispersed, unstable and highly customized excel sheets
 - Personal storage solutions
 - No description on data (Metadata)
 - Data processing steps and provenance not defined/tracked
 - No link, no structure, no context

Phenotyping experiments : Dealing with a big mess!

- ❖ Complex & massive data: source, type, operator, scale, transformations, etc.
 - Breeding, genomic, weather, soil, ... conditions, sensors, internet, etc
- ❖ Bad habits dealing with this mess
 - Dispersed, unstable and non-standardized excel sheets
 - Personal storage & management
 - No description on data (metadata)
 - Data processing steps and provenance not tracked
 - No link, no structure, no context

Need for action

Phenotyping data : No stress, we're going to make it!

Understand, adopt and apply FAIR principles

Findable

Metadata and data should be findable for both humans and computers

Interoperable

Data needs to work with applications or workflows for analysis, storage and processing

F

A

I

R

Accessible

Once found, users need to know how the data can be accessed

Reusable

The goal of FAIR is to optimise data reuse via comprehensive well-described metadata

<https://scibite.com/>

FAIR data

Data structure ⇒ store, retrieve, process data and Implement good practices

Based on two key elements:

Identification



- **Standardized & unambiguous identification**
 - Strategy and appropriate tools

Semantic



- **Based on Ontology:**
 - Data understanding (definition)
 - Data organization

FAIR data

MetaData in phenotyping experiments

Metadata is: Data 'reporting'

- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?



Photo by Michelle Chang. All Rights Reserved

What is Metadata



MetaData management in 6 levels

1: Description

No need to any special tool
⇒ Only human readable



description2014Bp45.txt

Id : <http://www.inrae.fr/PechRouche/2014Bp45>
Plot Beausoleil
Site = Pech Rouge
Position = « R4-P5 »
Carignan
The plot is supervising by Jean
planted : 2014

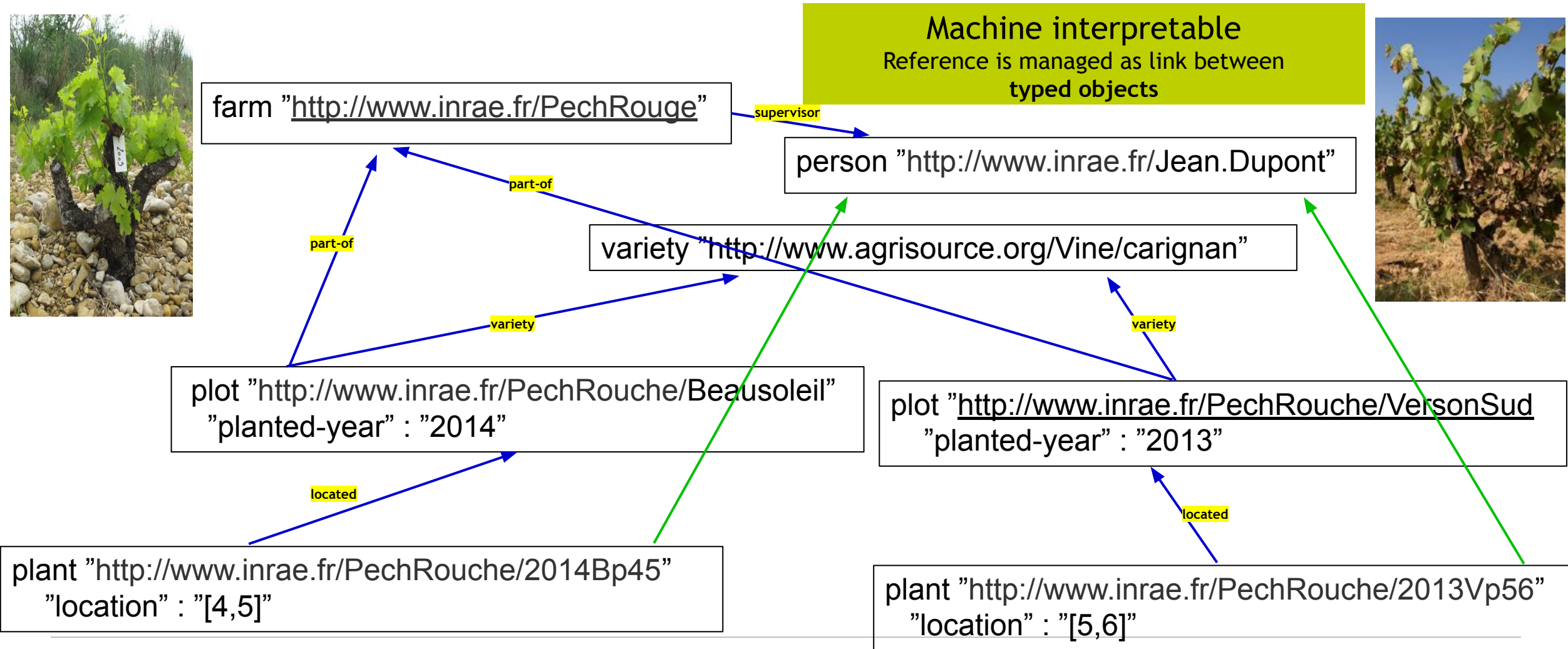


description2013Vp56.txt

ID : <http://www.inrae.fr/PechRouche/2013Vp56>
Verson Sud
Farm = Pech-Rouge
Location = [5,6]
Carignan
The plot manager Jean Dupont
planted : 2013

MetaData management in 6 levels

6: Description+Syntax+Vocabulary+Link+ Inference (reasoning)



MetaData management in 6 levels

6: Description+Syntax+Vocabulary+Link+ Inference (reasoning)

- Semantic resources
 - Ontology / Thesaurus
 - Rules
 - Standard term sets
- Specific tools
 - RDF (Subject-Predicate-Object)
 - Language (OWL / SWRL)
 - Graph DB systems
 - Reasons

Use standards



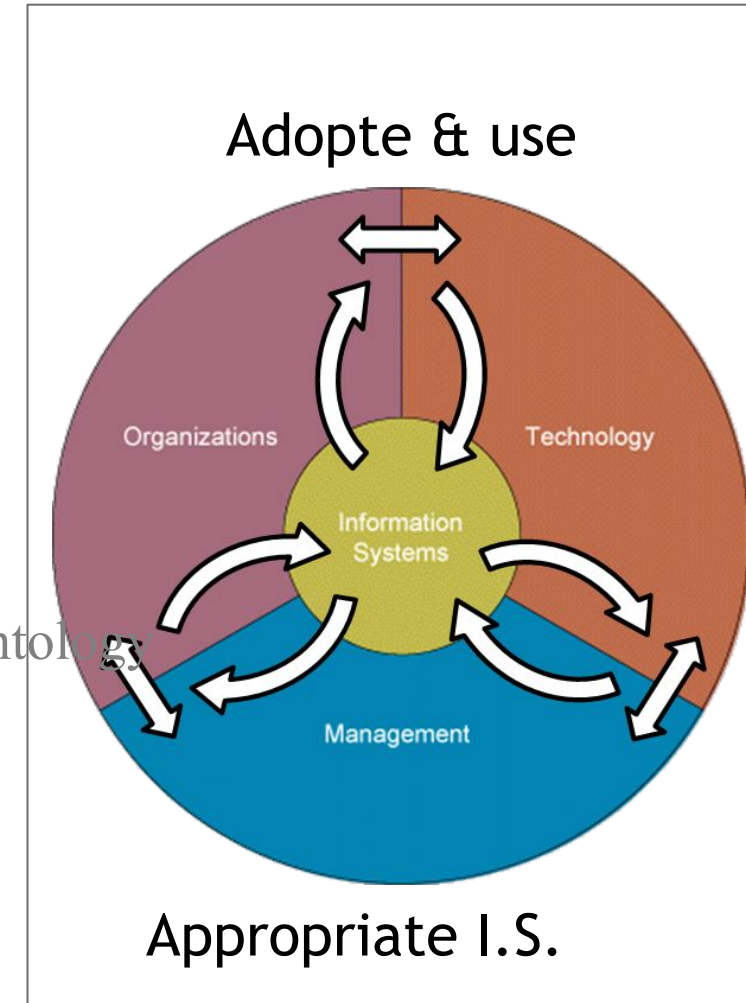
Common framework to share and reuse data

FAIR data: What do we need?

- Data base management system \Rightarrow SGBD, NoSQL, RDF storage
- Representation language \Rightarrow XML or JSON
- Use standards \Rightarrow MIAPPE, BrAPI
- Access (Web pages, Web Services) \Rightarrow BrAPI
- Knowledge representation (ontology, thesaurus, taxonomy): \Rightarrow Crop Ontology
- Analysis and visualisation tools (R, python,)

FAIR data: What do we need?

- Data base management system \Rightarrow SGBD, NoSQL, RDF storage
- Representation language \Rightarrow XML or JSON
- Use standards \Rightarrow MIAPPE, BrAPI
- Access (Web pages, Web Services) \Rightarrow BrAPI
- Knowledge representation (ontology, thesaurus, taxonomy): \Rightarrow Crop Ontology
- Analysis and visualisation tools (R, python,)



Overview

- Quick review of Session 1
- **Data Storage - Introduction**
- Cloud Computing
 - Data Storage Services
 - Data processing Services
- Data Management Policies
 - DMP - Data Management Plan
 - GDPR - General Data Protection Regulation

Data Storage - Introduction

Data Challenge

- ❖ **Context: more and more data!**
 - Cheap storage capacity and high speed network
 - e.g. **1 Gigabyte price** : \$400K in 1980, \$10K in 1990, \$10 in 2000, **now less than \$0.01**
 - Heterogeneous devices, simulations, machine learning, Internet data sources (Open, collaborative, etc) are available

Make data valuable!

- ❖ **Knowledge discovery**
- ❖ **Decision support**
- ❖ **Artificial Intelligence and machine learning**
(predict, detect, recognize, diagnostic, etc)

Share and Reuse Data

Data Storage - Introduction

Data Challenge

- ❖ **Volume:** massive data and exponential growing size
 - Hard to Store, Manage & Analyze
- ❖ **Variety, Vocabulary and complexity** (different sources, scales, disciplines, semantics, schema, format, etc)
 - Hard to understand, combine & Integrate
- ❖ **Velocity** : rate of data generation
 - Have to be processed on-line
- ❖ **Veracity, Validity, Variability**, Vulnerability, Volatility, Visibility, Visualisation, Vagueness, etc.
- ❖ **VALUE**

Data storage - Introduction

Phenomics data : Massive & shared data

! No personal computer

⇒ **Managing physical servers needs:**

- **Admin group with specific skills**
- **Flexibility** (variations in data production)
- **Specific rooms & facilities**

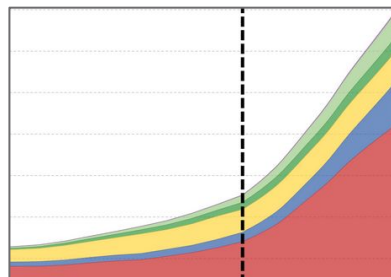


Data Storage - Introduction

Phenomics data : Massive & shared data

But exponential growth of data **requires continuous change** for physical servers:

- **a lot of effort for administrator group**
- **difficult because discontinuous resources**
- **and harder to manage**



Data Storage - Introduction

Phenomics data : Massive & shared data

! No p
E
A
E
S



Overview

- Quick review of Session 1
- Data Storage - Introduction
- Cloud Computing
 - Data Storage Services
 - Data processing Services
- Data Management Policies
 - DMP - Data Management Plan
 - GDPR - General Data Protection Regulation

Questions & Survey

Cloud Computing

On-demand: A consumer can provision (anticipate and use) computing capabilities (storage, VM, etc.) as needed automatically without requiring human interaction with a provider

Network access: Capabilities are available over the network and accessed through standard and use by light clients (phones, tablets, laptops, etc.)

Resource pooling and elasticity: resources are dynamically assigned and reassigned according to consumer demand.

Capabilities can be elastically provisioned and released to scale rapidly with demand.

To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time

Measured service: Cloud systems automatically control and optimize resource use. Resource usage can be monitored, controlled, and reported



Cloud Computing

Relevant services for data management

- **Storage**
- **Backup, Archiving**
- **Virtual Machine**
- **High Performance Computing (HPC), High Throughput Computing (HTC)**
- **DataHub**
- **Virtual Research Environment**
- ❖ **Coordination of cloud services for research: EOSC**
 - Academic European providers: EGI, EUDAT**

Cloud Computing




EOSC: European Open Sciences Cloud

The screenshot shows the EOSC website interface for researchers. At the top, there are three navigation tabs: "For Researchers" (selected), "For Providers", and "For Businesses". Below the tabs, the text reads "Researchers including scientists, students, lecturers, teachers and citizen scientists". The main content is organized into three columns: "Explore and Contribute", "Tools", and "More".




For Researchers | **For Providers** | **For Businesses**

Researchers including scientists, students, lecturers, teachers and citizen scientists

Explore and Contribute

-  [Discover Research Outputs](#)
Find datasets, scientific publications and software for your research activities
-  [Publish Research Outputs](#)
Store, backup, archive your data, publications, software
-  [Find Funding Opportunities](#)
Learn about RDA/EOSC Future open calls, EOSC DIH support schemes and more

Tools

-  [Access Computing and Storage Resource](#)
Find HPC, IT centres for science, cloud computing, online storage
-  [Process and Analyse](#)
Verify, organise, transform and integrate data, then export it in the format you need
-  [Access Training Materials](#)
Find lessons, courses, videos

More

- [Research Data Management](#)
- [Research Infrastructures](#)
- [Instruments & Equipments](#)
- [Regional & Thematic Projects](#)

[Get Inspired](#)

Cloud Computing

EOSC: Data management services

Filters

Research step [clear](#)

- Discover Research Outputs (5493877)
- Process and Analyse (84)
- Manage Research Data (51)
- Access Training Material (36)
- Access Computing and Storage Resources (32)
- Access Research Infrastructures (25)
- Publish Research Outputs (21)
- Find Bundles (8)
- Find Instruments & Equipment (1)

51 search results All catalogs

[Clear filters](#) Research step: Manage Research Data [×](#)

<< < **1** 2 3 4 5 >

Sort By **Best match** [▼](#)



B2SAFE

Data Source

ORDER REQUIRED

Horizontal Service

↓ 0 Downloads 204 Views English

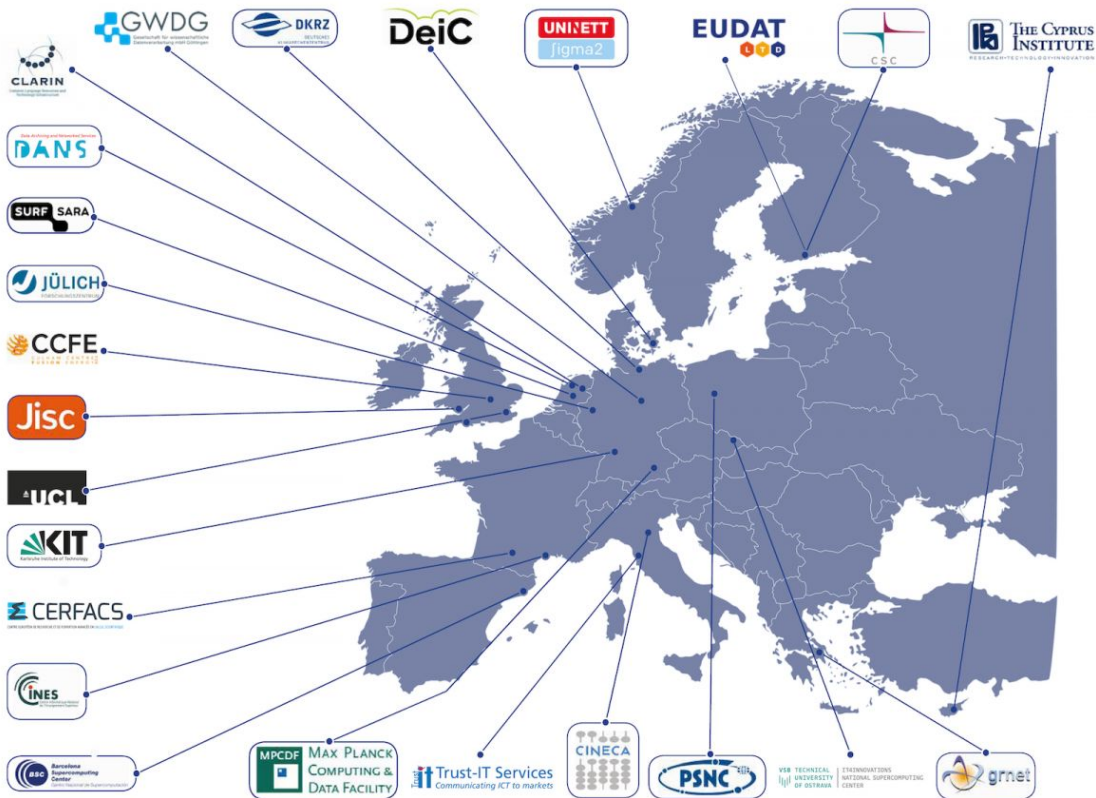
Organisation: **EUDAT**







Scientific domain: **Generic**

B2SAFE is a robust and highly available service which allows community and departmental repositories to implement data management policies on their research data across multiple

Cloud Computing

EUDAT services



 B2ACCESS	 B2DROP	 B2FIND
B2ACCESS Identity & authorisation View service	B2DROP Sync and share research data View service	B2FIND Find research data, research data portal View service
 B2HANDLE	 B2SAFE	 B2SHARE
B2HANDLE Register your research data with a persistent identifier View service	B2SAFE Keep research data safe via data management policies View service	B2SHARE Store and publish research data View service

Cloud Computing

EGI services



Explore our solutions

By Use Case

1 Batch computing

Enables researchers, and scientific communities to easily and efficiently run hundreds of thousands of batch computing jobs on the EGI Infrastructure

[Discover more](#)

2 Interactive computing

Web-based environment to facilitate the sharing and reproducibility of Open Science

[Discover more](#)

3 Federated access

Identity and access management solution to increase productivity and secure access to services and resources

[Discover more](#)

4 Data Space

Provisioning of integrated compute and data capacity, data collections, as well as cloud-enabled service offering for scalable data analysis

[Discover more](#)

5 Data Federation

Unified data access across distributed data providers

[Discover more](#)

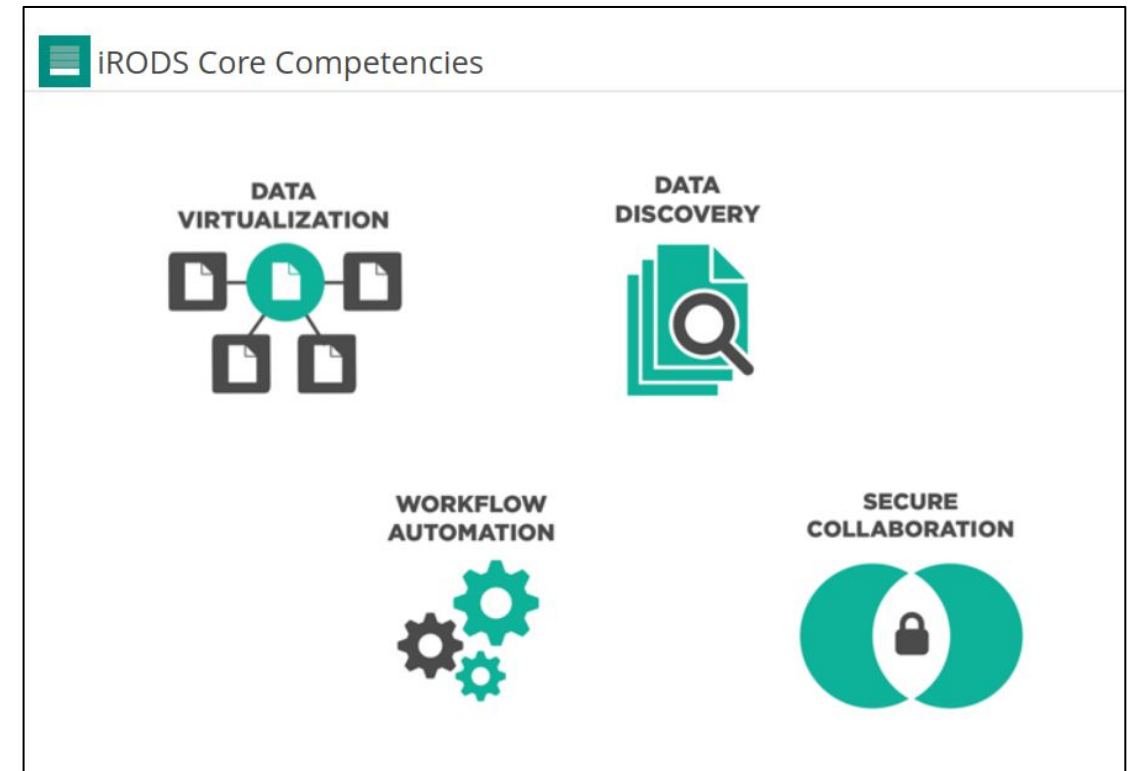
6 Service hosting

Dedicated and secure servers to deploy and scale-up domain-specific web hosting solutions on the EGI Federated Cloud

Storage Service

Technological solutions for data storage: iRODS, OneDATA, S3

- Access - Authentication, Authorization, Revocation
- Description - Standards for discovery, compliance
- Integrity - Confidence that nothing has changed
- Replication - Multiple copies, multiple locations
- Availability - If things are down, nothing else matters
- Migration - Hardware changes, format changes
- Recovery - Robust plans for when things go wrong
- Provenance - Full record of all related activity
- Retention - Deleting data on a defined schedule



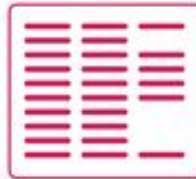
Storage Service

Based on metadata templates

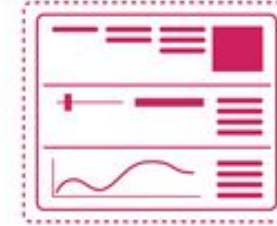
USER-FRIENDLY
FORMATTING



iRODS AVUs

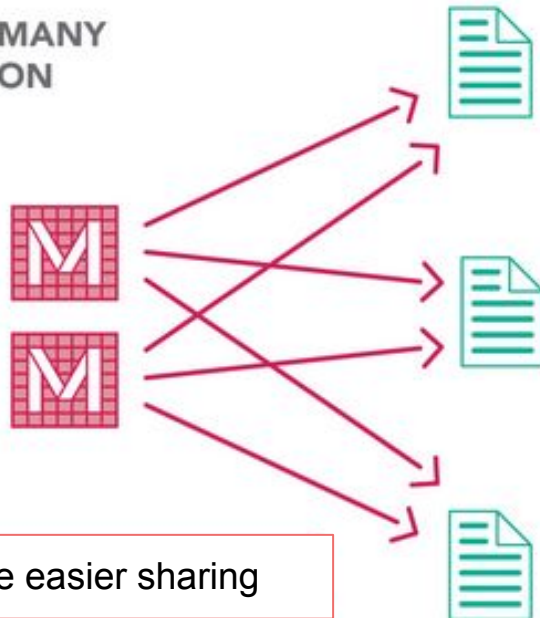


AVUs VIEWED USING
A METADATA TEMPLATE



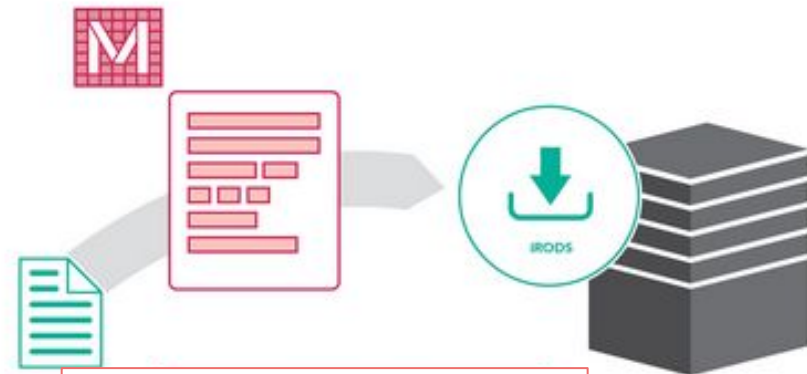
Monitoring

MANY-TO-MANY
APPLICATION



Make easier sharing

VALIDATION AND ENFORCEMENT



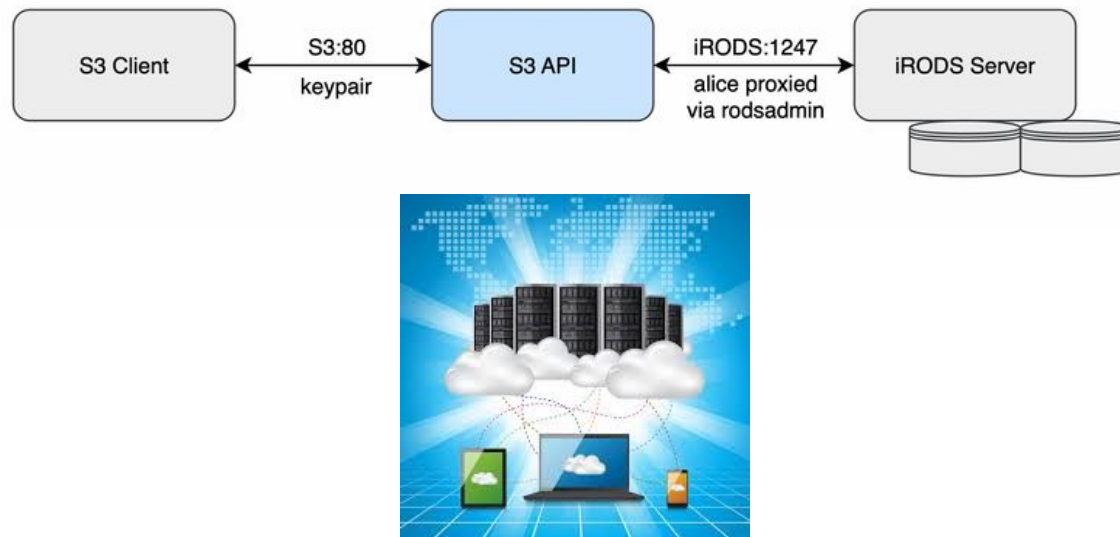
Metadata set allow to find
and validate datafile



Storage Service

Technological solutions: iRODS, OneDATA, S3

- Tools are available for the interoperability between technologies



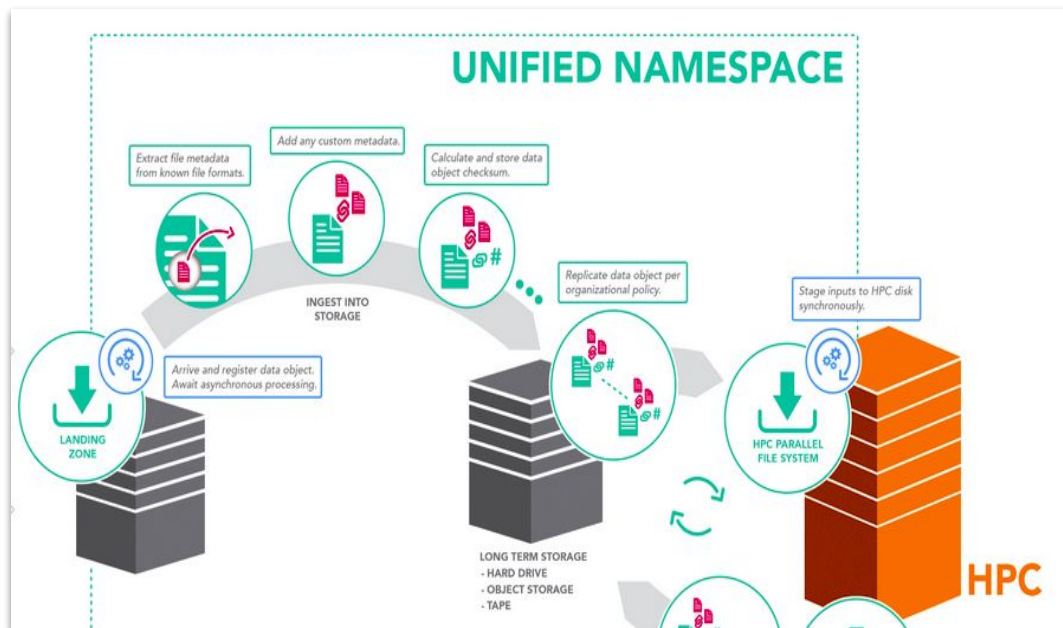
Implementation and use via API, Web interfaces, virtual hard disk

Questions ?

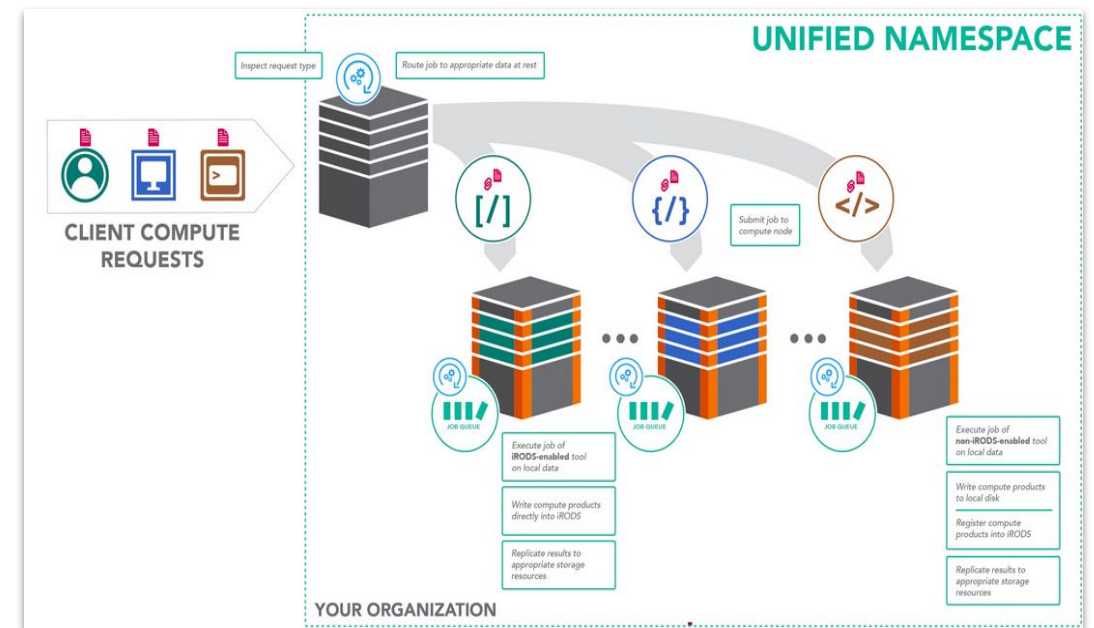
Storage & Computing Considerations

Multiples data access patterns according different usages

- **Data to compute (classic approach)** : Transfer data to computes nodes and execute tasks
- **Compute to data** : Distributes and execute computing tasks near the data nodes



Data to compute



Compute to data

Computing Services

HPC -> rather focus complex computing



HTC -> rather focus a large number a data
(Data-Based same operation on different data)



Computing Services

HPC	HTC
High-Performance Computing	High Throughput Computing
use of multiple computer processors in order to perform complex computations parallelly.	parallelly executes a large number of simple and computationally independent tasks .
running large-scale, complex , and computationally intensive applications that need significant resources and memory.	running a large number of tasks that does not require a large amount of memory and resources.
designed to provide speed for large tasks.	designed to increase the number of tasks
Centralized management	Distributed management
Try to reduce the risk of data loss and data corruption	Do not affect any other running processes.
Few users are running together	Scale horizontally for simple tasks .
Complex decision support, weather forecasting .	Bioinformatics or Phenomics.

Questions ?

Datahub Service

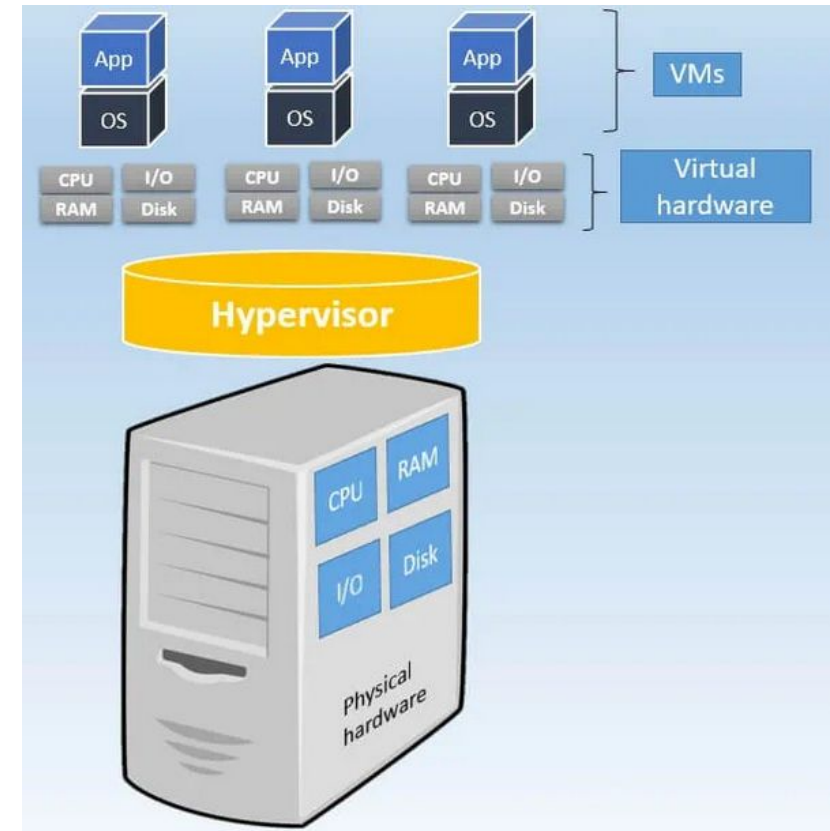
A datahub homogenizing data and possibly serving data in multiple desired formats with features such as de-duplication, quality, security, and standardized set of query services. Datahub is more structured than data lake.

- Unique (virtually) data source



Virtual Machine

- **Virtual machine (VM)** is the emulation of computer systems
- Provides same functionalities as physical machine
- Several virtual machines on one physical machines
- Can run different Operating Systems
- Hardware abstraction



Virtual Machine

Physical services	Virtual Machines
Large upfront costs	Small upfront costs
Physical servers and additional equipment take a lot of space	A single physical server can host multiple VMs , thus saving space
Has a short life-cycle	Supports legacy applications
No on-demand scalability	On-demand scalability
Hardware upgrades are difficult to implement and can lead to considerable downtime	Hardware upgrades are easier to implement; the workload can be migrated to a backup site for the repair period to minimize downtime
Difficult to move or copy	Easy to move or copy
Poor capacity optimization	Advanced capacity optimization is enabled by load balancing
Doesn't require any overhead layer	Overhead is required for running VMs
Running stable services and operations which require highly productive computing hardware	Running multiple services for multiple users, which plan to extend in the future

Backup Service

Insurance against human error, technical failure, disaster

Protect your current work

Incremental (difference management)

Overwritten

Restore process



Archive Service

Build an historical record

Valuable content to keep

Permanent record

Eco friendly (low energy)



Backup vs Archive

	Backup	Archive
Definition	COPY -> PASTE	CUT -> PASTE
Purpose	Disaster Recovery Accidental Data Loss Compliance	Storage Cost Compliance Analytics
Performance	Significantly drops the performance	Drastically improve the performance
Retention	Short-term	Long-term
Cost	More costly day by day	Significantly save storage cost

Questions ?

Data Infrastructures

Data Challenge

- ❖ **Volume:** massive data and exponential growing size
 - Hard to Store, Manage & Analyze
- ❖ **Variety, Vocabulary and complexity** (different sources, scales, disciplines, semantics, schema, format, etc)
 - Hard to understand, combine & Integrate
- ❖ **Velocity** : rate of data generation
 - Have to be processed on-line
- ❖ **Veracity, Validity, Variability, Vulnerability, Volatility, Visibility, Visualisation, Vagueness, etc.**
- ❖ **VALUE**

Data Management Challenge : Velocity

- Green House platforms produce tens of thousands images/day \Rightarrow (200 days/year)
- Field platforms produce tens of thousands images/day \Rightarrow (100 days/year)
- Omic platforms produce tens of Gbytes/day \Rightarrow (300 days/year)

❖ Approaches:

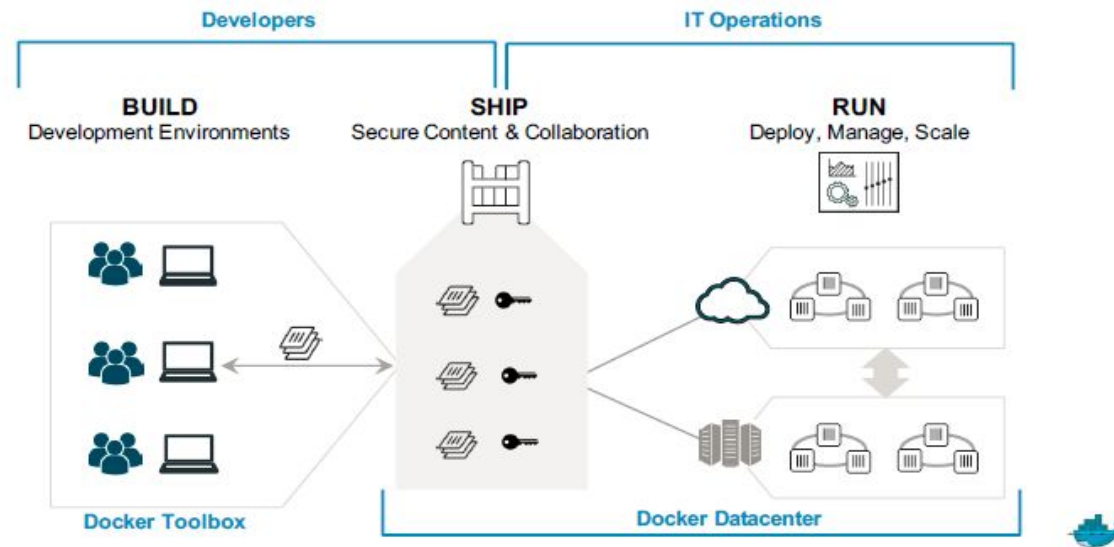
- **Scientific Workflow** (consists of an orchestrated and repeatable pattern of tasks of information processing (depicted as a sequence of operations)
 - **Galaxy**
 - **Cloud workflow services**
- **Allows to share and to store data treatment**

Container Service

Container : Code and dependencies (libraries, DB) and configuration inside a single package

Container service:

- Provide an **easy-to-use** and **reproducible** execution environment
- **Simplified integration, deploy and management** of applications
- Deploy **scalable and secure** applications
- Allow developers to **focus on user features**, not on infrastructure management



VRE Service

A **virtual research environment (VRE)** or **virtual laboratory** is an online system for **research collaboration**

Can be based on virtual desktop

Standard features:

Forum, Data Publication

Data exploration, Data visualisation

Set of tools for data analytics

Set of discipline-specific tools



Virtual Research Environment

to access, share and collaborate



Data storage - Conclusion

- On-demand infrastructure and Elasticity
- Virtualization technologies
- Many services available: complementaries with overlaps
- **Coordination of metadata and data services requires DMP**

- Quick review of Session 1
- Data Storage - Introduction
- Cloud Computing
 - Data Storage Services
 - Data processing Services
- Data Management Policies
 - DMP - Data Management Plan
 - GDPR - General Data Protection Regulation

Questions & survey

How to formalise your data management?

In a research project or structure

You need to describe:

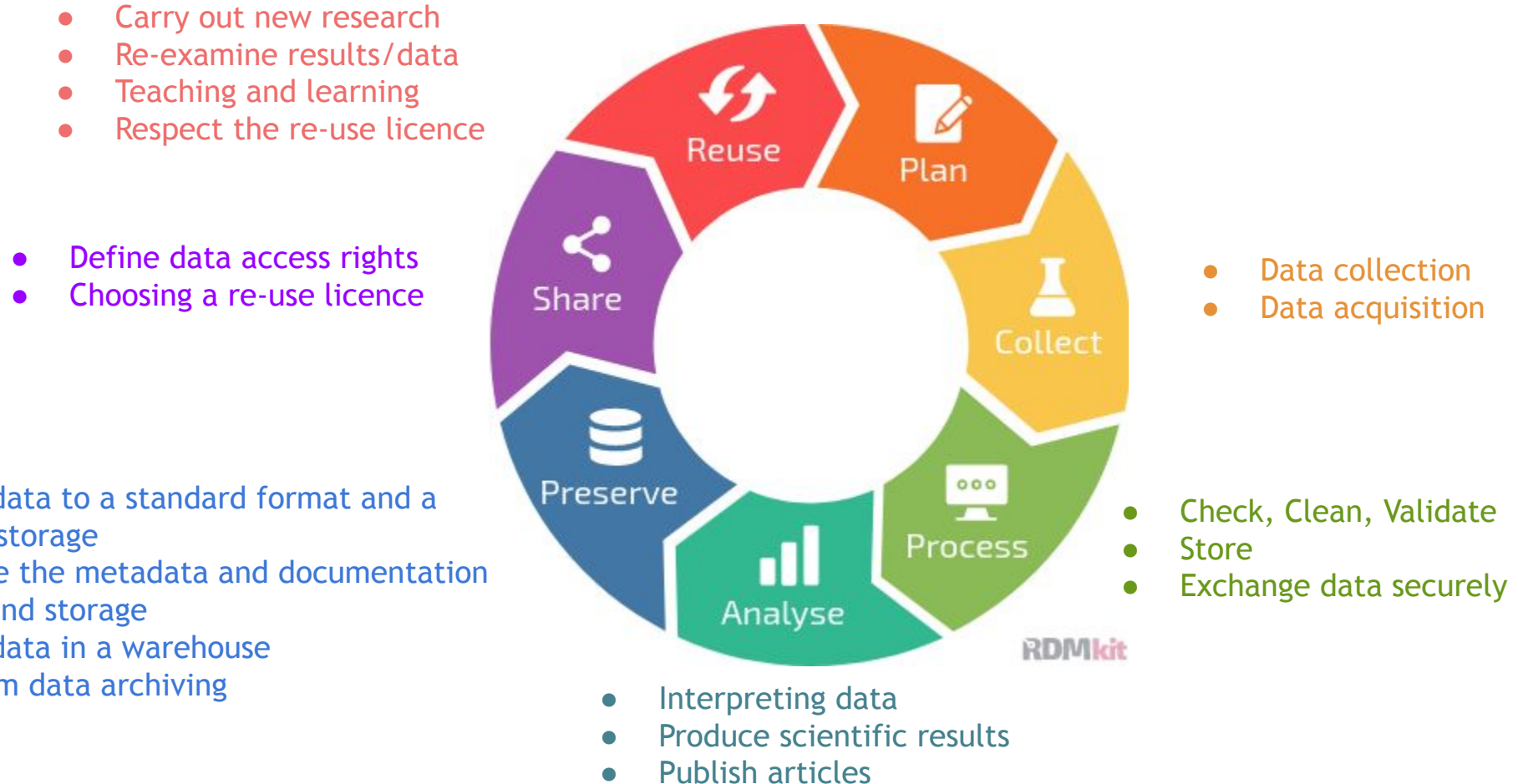
- Various sources and types of data
- Various processing steps applied to data
- Storage resources and duration
- Data protection, rights of use

⇒ That's why we need **Data Management Plan!**

DMP - Data Management Plan

What is a Data Management Plan (DMP)?

A DMP is a document that describes how the data of a research project or structure will be managed throughout its lifecycle.



DMP - Data Management Plan

What is it for?

Implementing best practice, respecting the FAIR principles

- Ensuring the reproducibility of experiments by describing the data and how it was obtained
- Enabling data to be understood and therefore re-used
- Avoid loss of data by appropriate storage
- Establish the roles and responsibilities of each party
- Respect the law and individuals by clarifying the legal and ethical framework (link to GDPR)
- Clarify re-use rights and sharing arrangements

DMP - Data Management Plan

How to set it up?

- **A pragmatic approach**
 - Simple to understand, implement, evaluate and develop
 - Text document for the time being, but "actionable machine" versions are appearing
- **Models available**
 - Beware of **national specificities!**
 - Frameworks defined by supervisory bodies (National Institutes, Universities, etc.), funders (HORIZON 2020, etc.), computing/storage centres (IN2P3), etc.
 - More or less rich information (INRAE framework = 40 questions, ANR framework = 15 questions, H2020 framework = 9 questions)
 - Little or no specificity about the type of data
 - The format and proposed aids change, but the content remains the same!
- **Tools to create DMP:**
 - DSW (Data Stewardship Wizard): <https://ds-wizard.org>
 - B2SAFE: <https://www.eudat.eu/b2safe>
 - French tool - DMP OPIDoR: <https://dmp.opidor.fr>
 - RDMkit - https://rdmkit.elixir-europe.org/data_life_cycle

DMP - Data Management Plan

Project DMP vs. research structure DMP

Project DMP

- Funded or unfunded research project
- Specific scope and fixed duration
- Mandatory

Research structure DMP

- Research platform, Collective Scientific Infrastructure (EU, CRB, Platforms, etc.), Research Infrastructure
- Broader scope and indefinite duration
- Not compulsory, voluntary for the moment
- Distancing practices
- Identification of areas for improvement
- Formalisation of our requirements and commitments: linked to quality approach
- Centralisation of information
- Facilitate the creation of other DMP

DMP - Data Management Plan

Project DMP vs. research structure DMP

Project DMP

- Funded or unfunded research project
- Specific scope and fixed duration
- Mandatory

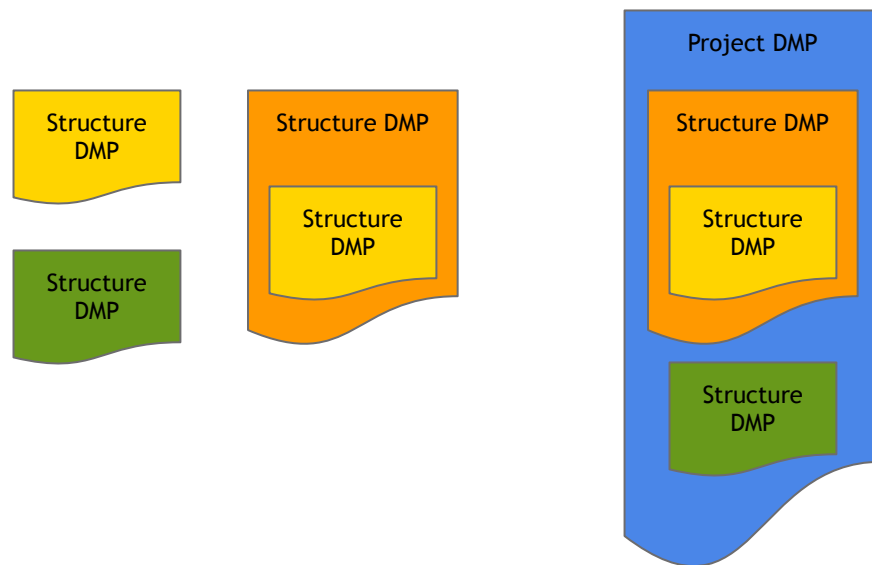
is based on



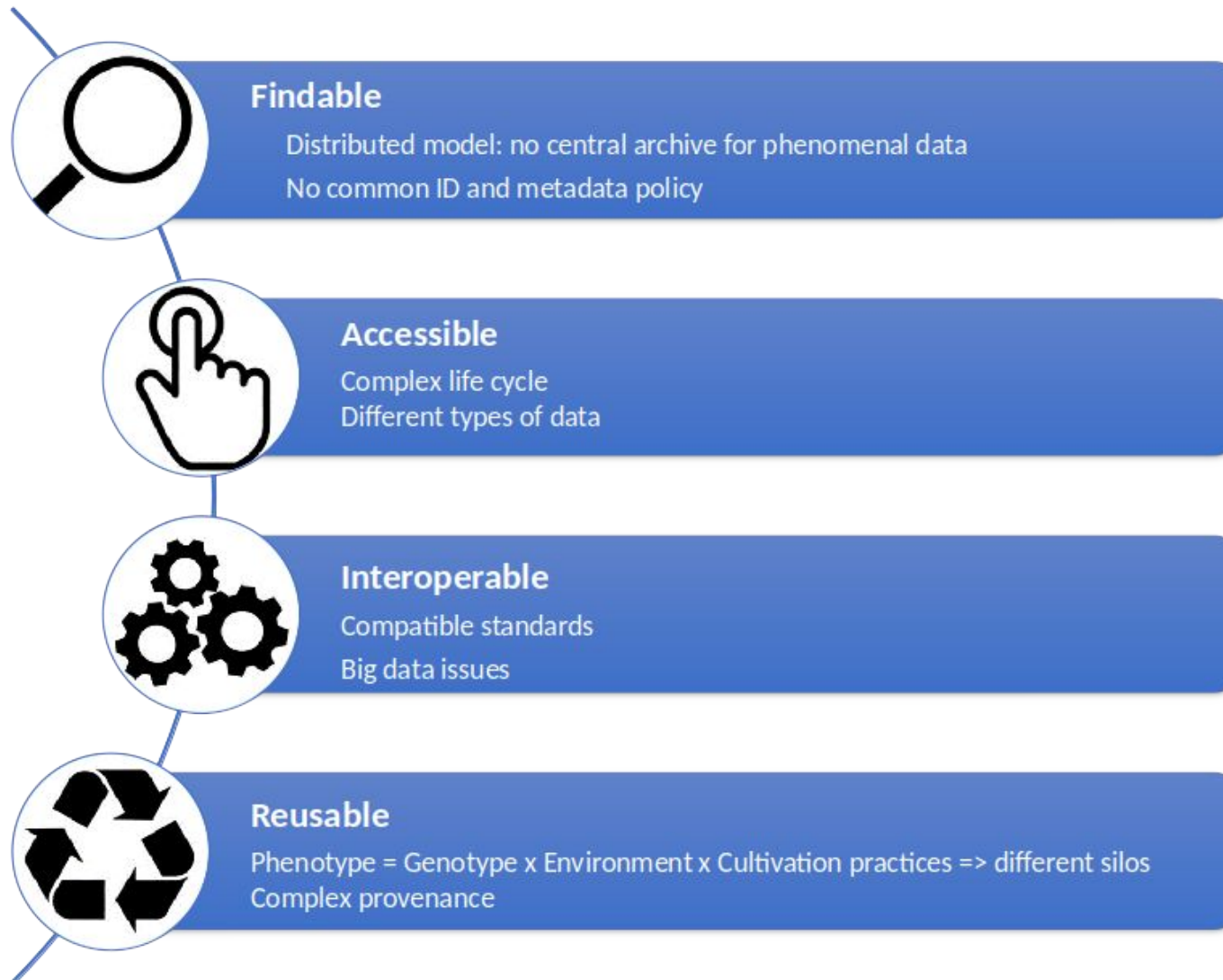
help with setting up

Research structure DMP

- Research platform, Collective Scientific Infrastructure (EU, CRB, Platforms, etc.), Research Infrastructure
- Broader scope and indefinite duration
- Not compulsory, voluntary for the moment
- Distancing practices
- Identification of areas for improvement
- Formalisation of our requirements and commitments: linked to quality approach
- Centralisation of information
- Facilitate the creation of other DMP



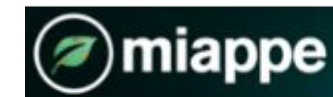
DMP - Phenotyping data issues



Data portals



Data Archives



Platform IS



Questions ?

Personal Data: a specific case

What is General Data Protection Regulation (GDPR)

- **G** - General
- **D** - Data
- **P** - Protection
- **R** - Regulation

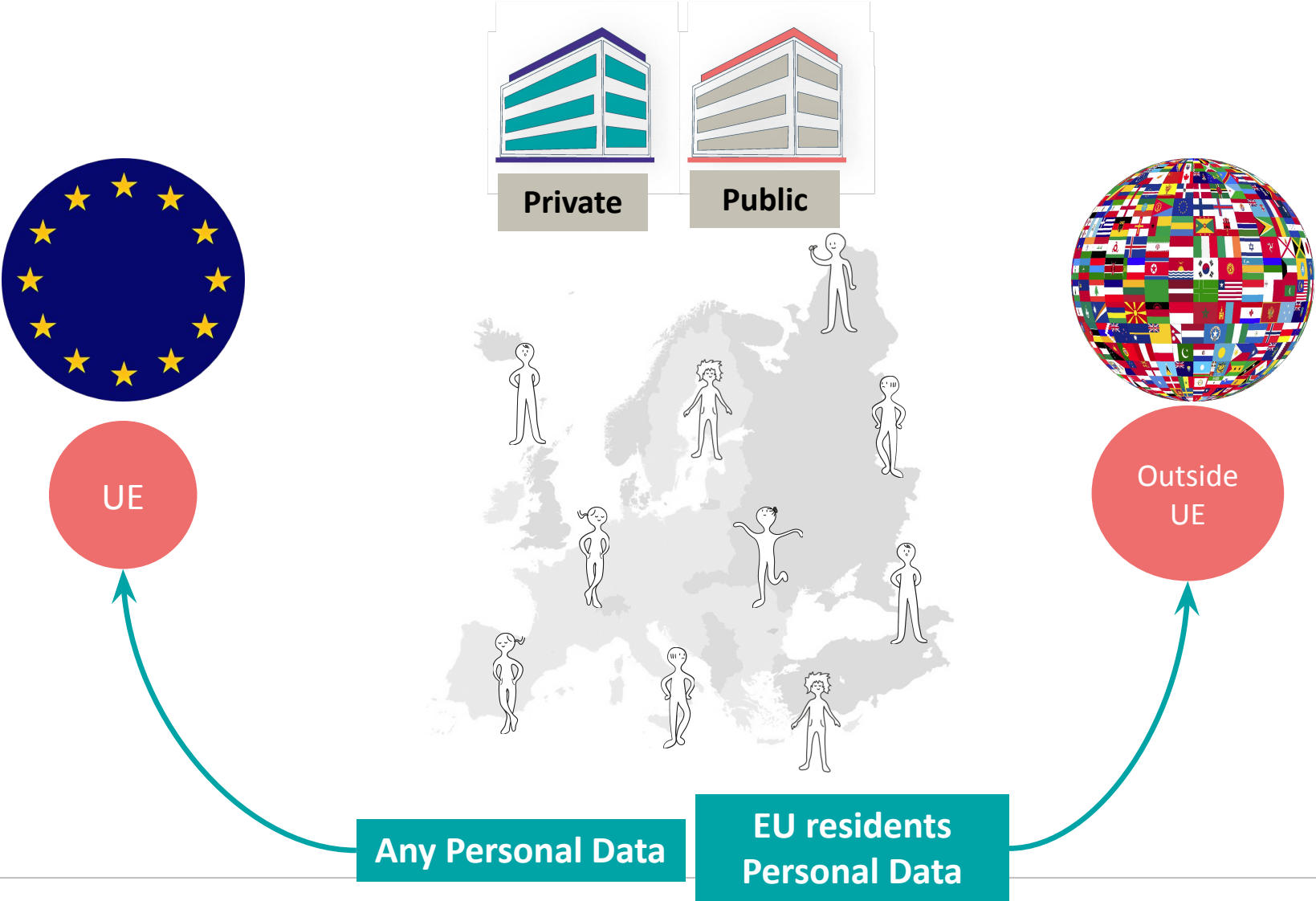
European Union law implemented **25th May 2018**

European reference text on personal data protection for EU residents.

GDPR - General Data Protection Regulation

Who is involved?

Any organization that processes the personal data of people in the EU must comply with the GDPR.



GDPR - General Data Protection Regulation

EU National Data Protection Authorities (DPA)

- DPAs are **independent public authorities** that supervise, through investigative and corrective powers, the application of the data protection law.
- As in the case of the DMP, the implementation of the GDPR has **specific national characteristics**
- DPA provide expert advice on data protection issues and handle complaints lodged against violations of the General Data Protection Regulation **and the relevant national laws.**
- There is **one DPA in each EU Member State** :
https://edpb.europa.eu/about-edpb/about-edpb/members_en

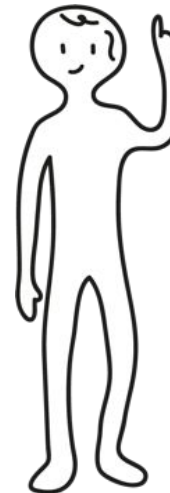
GDPR - Definitions & key concepts

Data Protection Officer (DPO)

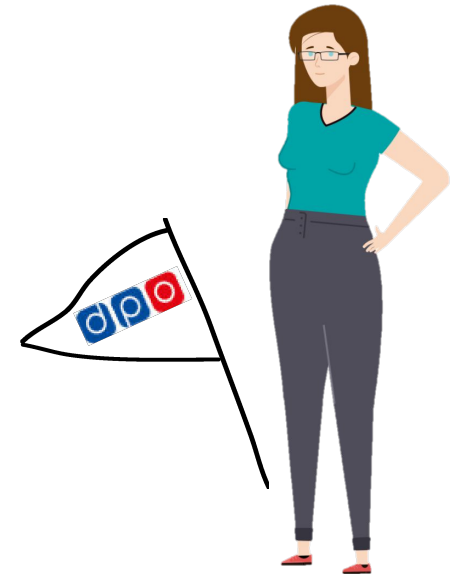
- DPO handles personal data and GDPR compliance
- DPA must be informed and have easy access to DPO contact details
- DPO must be provided with the **necessary resources** to carry out his duties effectively and independently.
- “The controller and the processor shall ensure that the data protection officer is involved, properly and in a timely manner, in all issues which relate to the protection of personal data.” (*GDPR Article 38*)

Compliance
file

I'm compliant
and I can prove it!



Advice

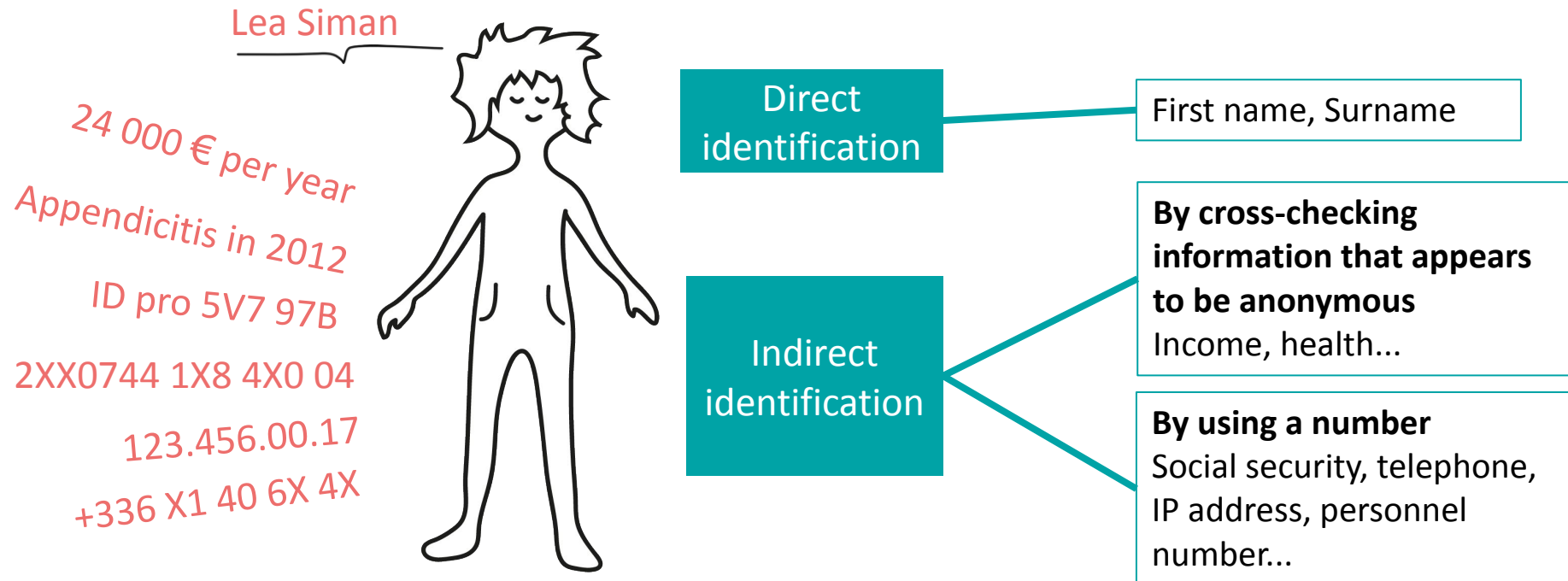


60

GDPR - Definitions & key concepts

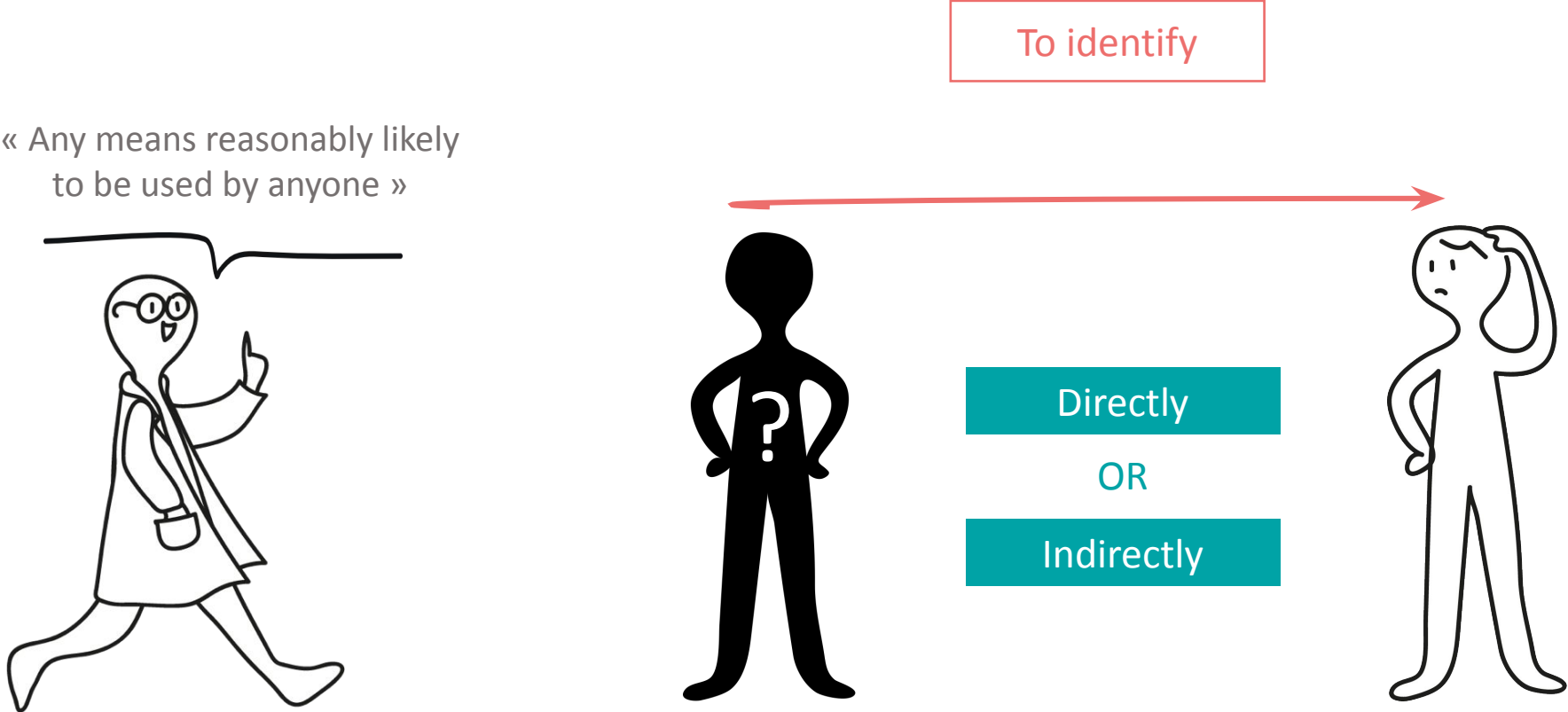
Personal data (PD)

- **Personal data** – Personal data is any information that relates to an individual who can be directly or indirectly identified. Names and email addresses are obviously personal data. Location information, ethnicity, gender, biometric data, religious beliefs, web cookies, and political opinions can also be personal data. Pseudonymous data can also fall under the definition if it's relatively easy to ID someone from it.



GDPR - Definitions & key concepts

Personal data (PD) - Determining whether a person is identifiable

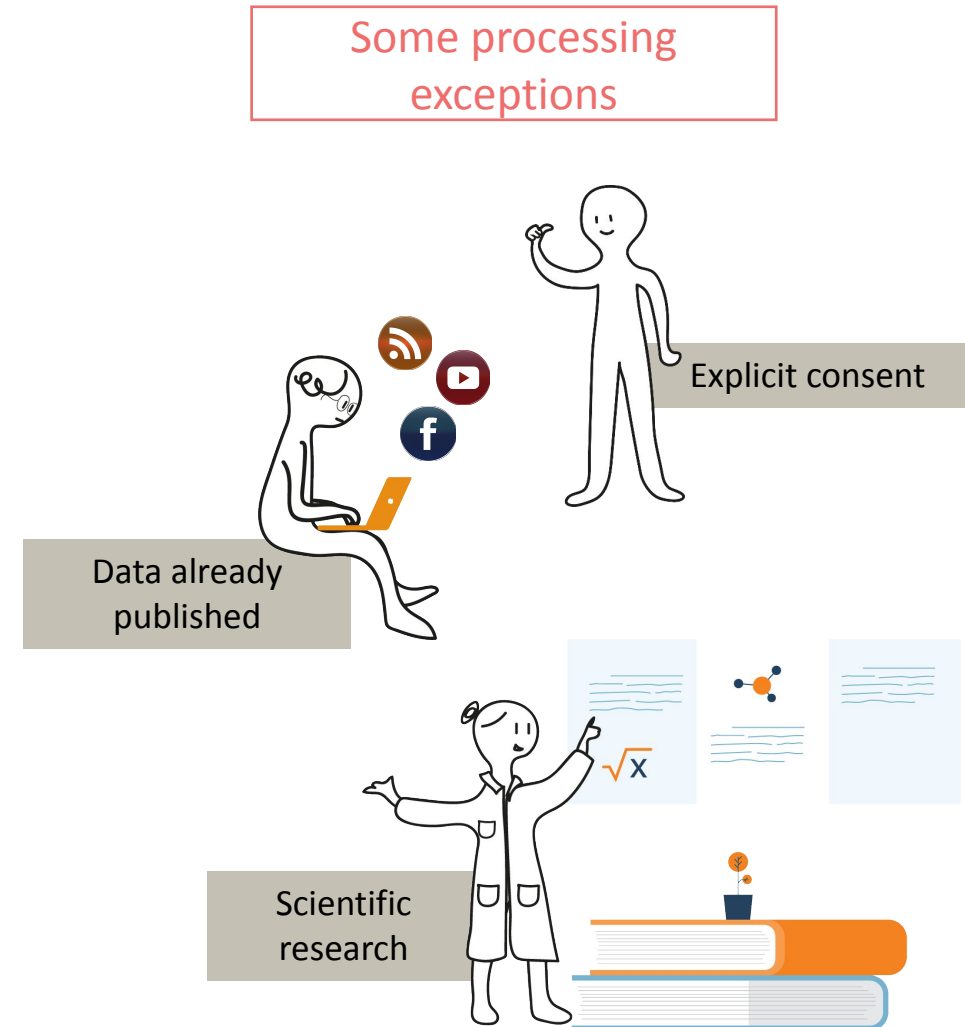


GDPR - Definitions & key concepts

Personal data (PD) - Sensitive data

Prohibited from processing

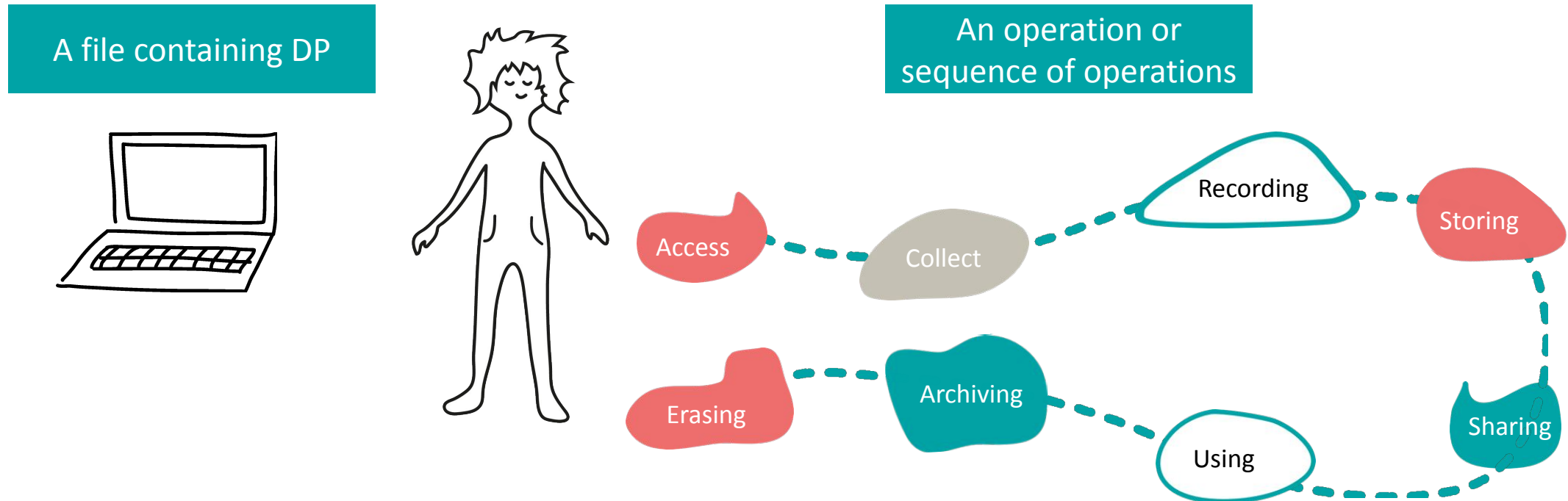
- Racial or ethnic origin
- Political opinions
- Philosophical or religious beliefs
- Trade union membership
- Health
- Sex life
- Genetic and biometric data
(for identification purposes)



GDPR - Definitions & key concepts

Data processing

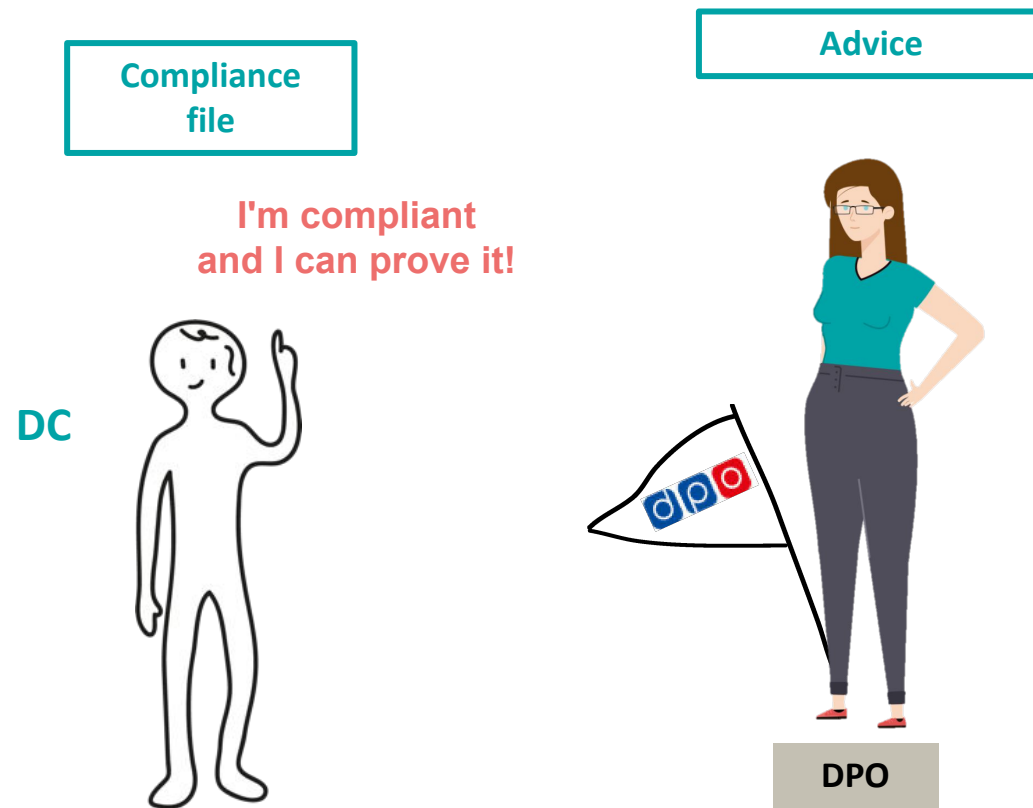
- **Data processing** – Any action performed on data, whether automated or manual. The examples cited in the text include collecting, recording, organizing, structuring, storing, using, erasing... so basically anything.



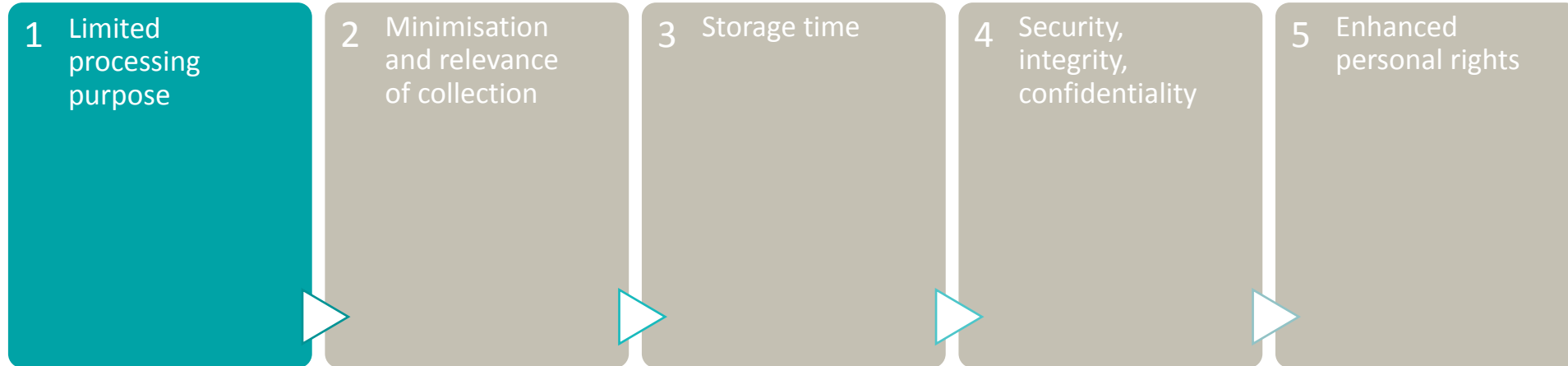
GDPR - Definitions & key concepts

Data controller (DC)

- **Data controller** – The person who decides why and how personal data will be processed. If you're an owner or employee in your organization who handles data, this is you.



5 Personal Data protection principles



PD collected for a specific and lawful purpose



Fair and transparent processing



Unchanged initial purpose

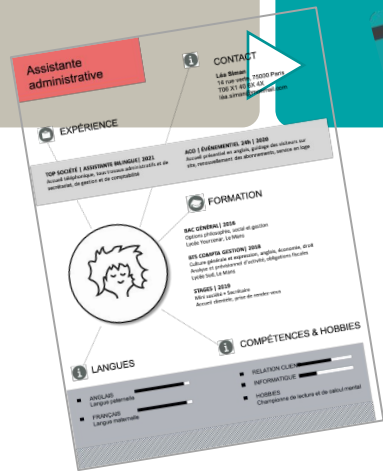
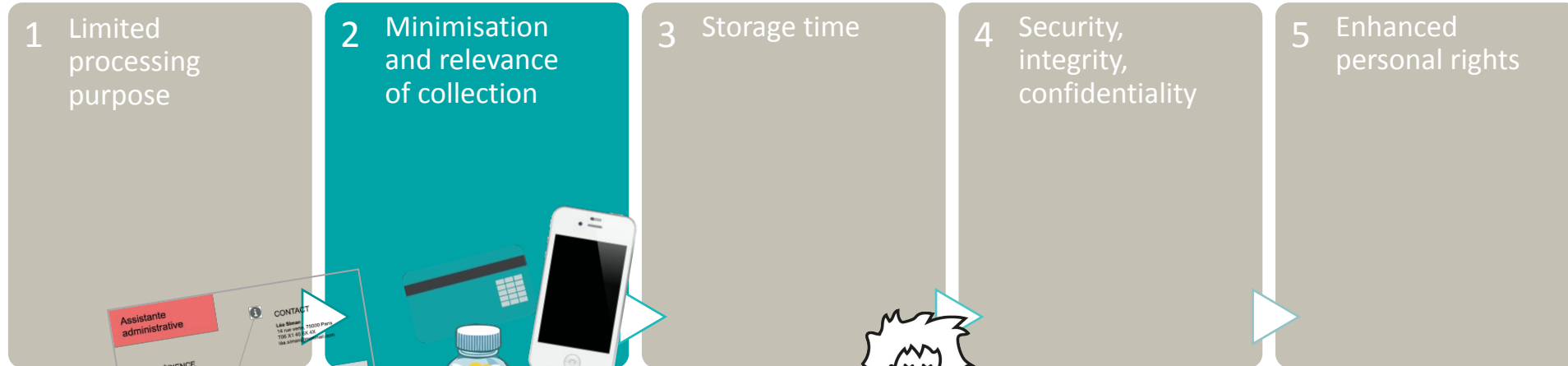
If further processing



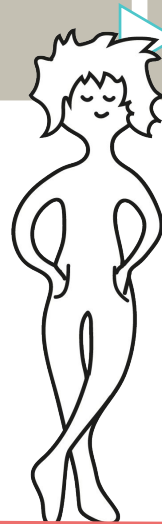
Exception to further processing: archives in the public interest, historical or **scientific research** or statistical purposes



5 Personal Data protection principles



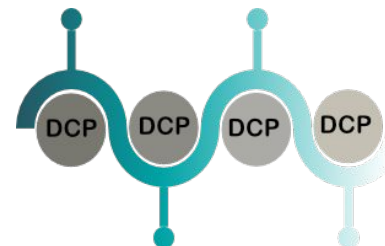
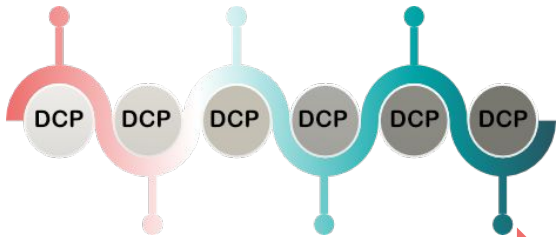
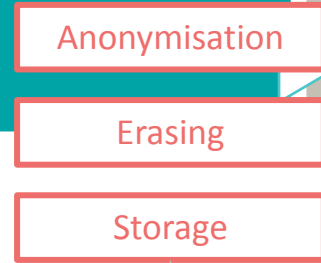
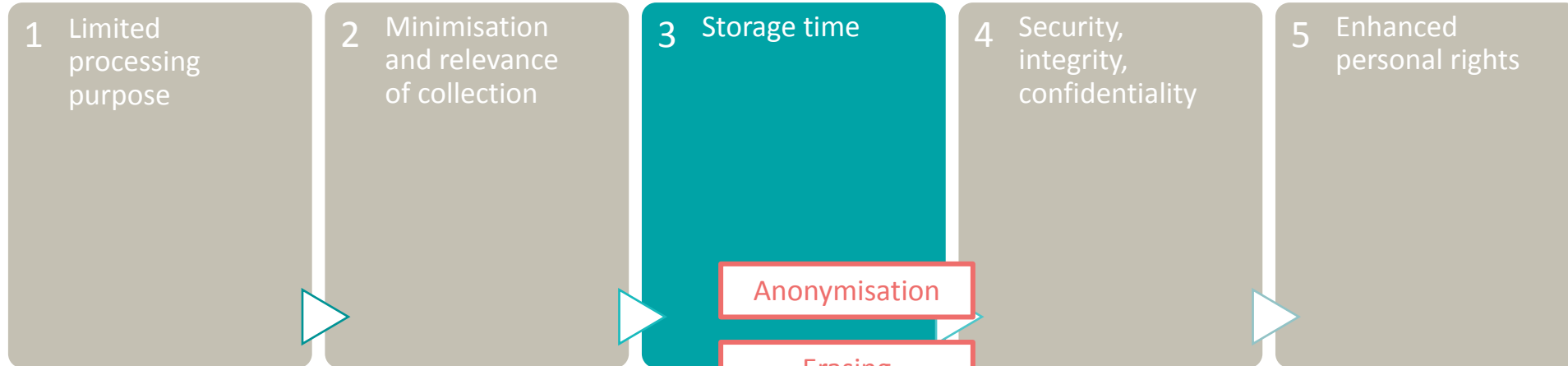
28K€ per year
ID pro 5V7 97B
2XX0744 1X8 4X0 04
Date of birth: 19/12/1985



Relevant and necessary DP collected

What are the purposes of data processing?

5 Personal Data protection principles



Data collecting

Data retention period and criteria

End of processing

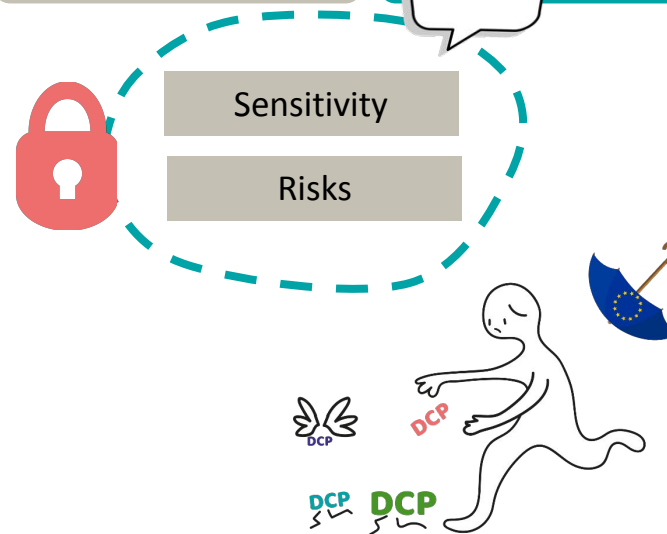
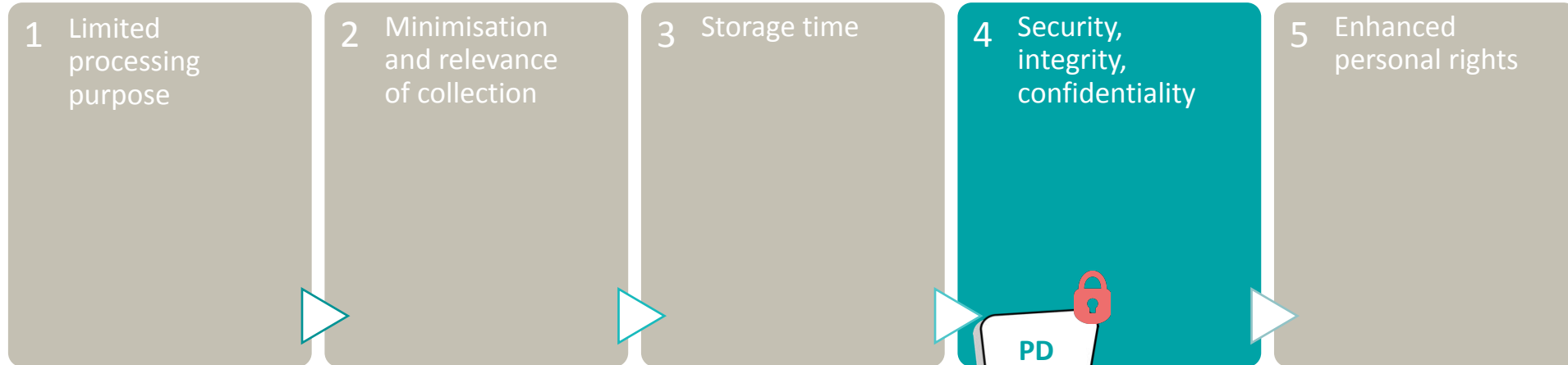
Extension of duration

Exception:
scientific
research

Mandatory security
to guarantee
people's rights



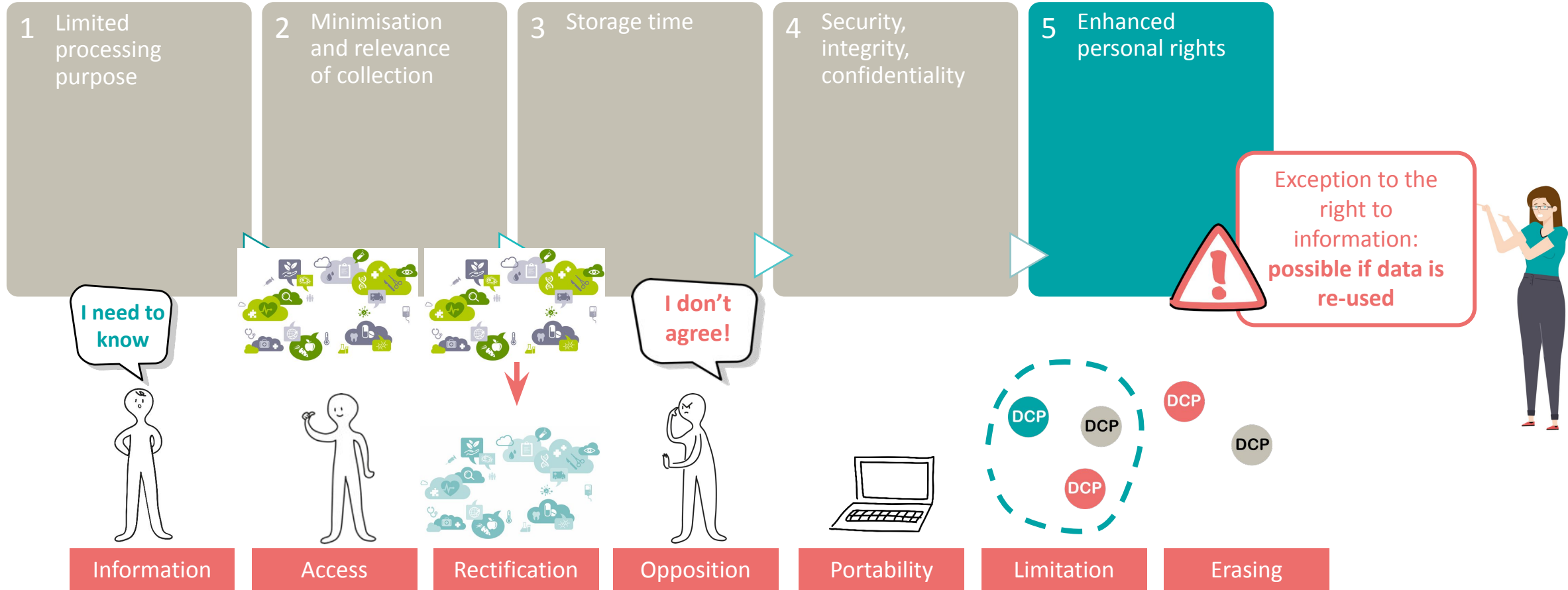
5 Personal Data protection principles



-
- Unauthorised use or access (Confidentiality)
 - Unwanted modification (Integrity)
 - Loss (Availability)

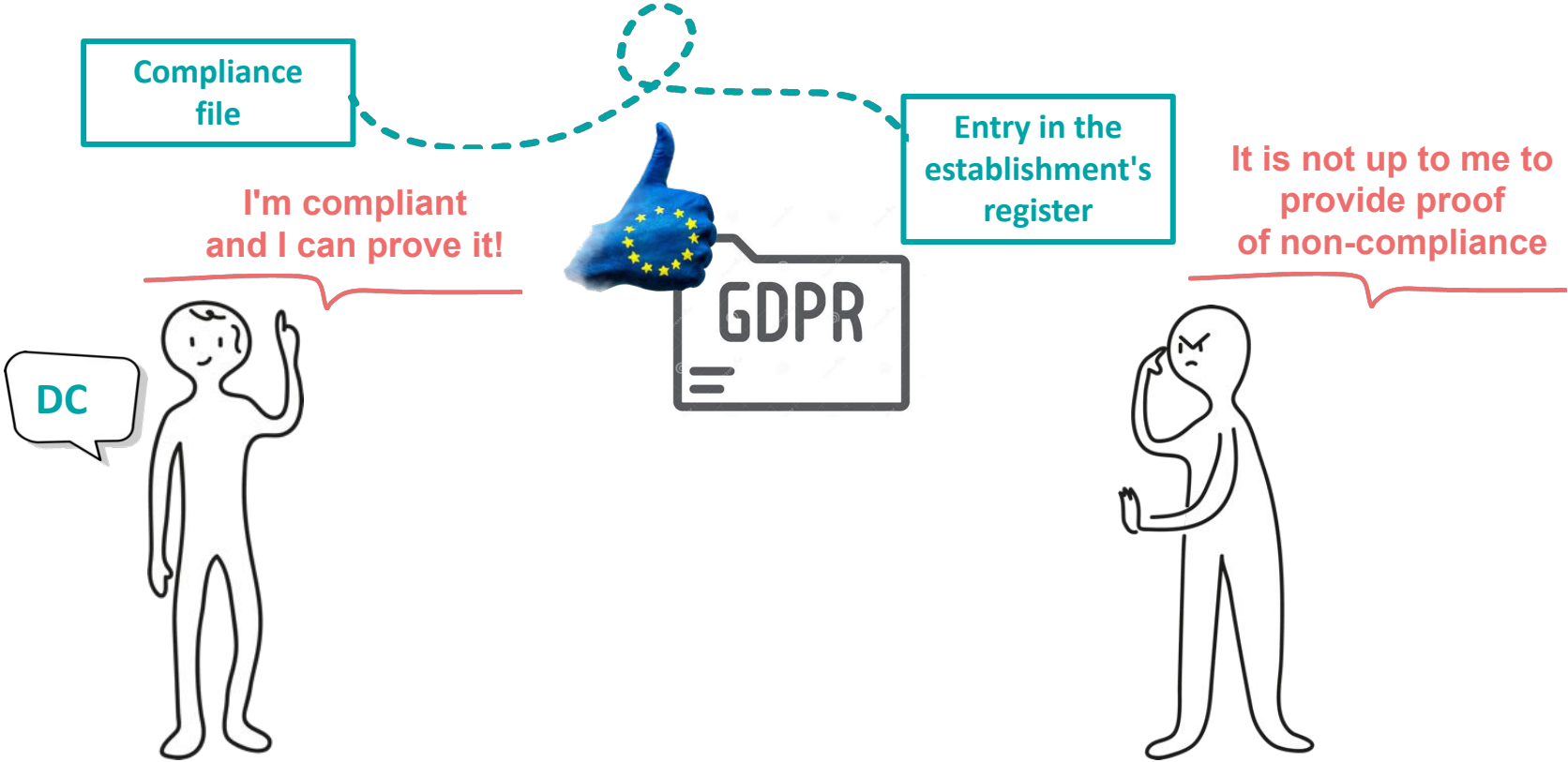


5 Personal Data protection principles



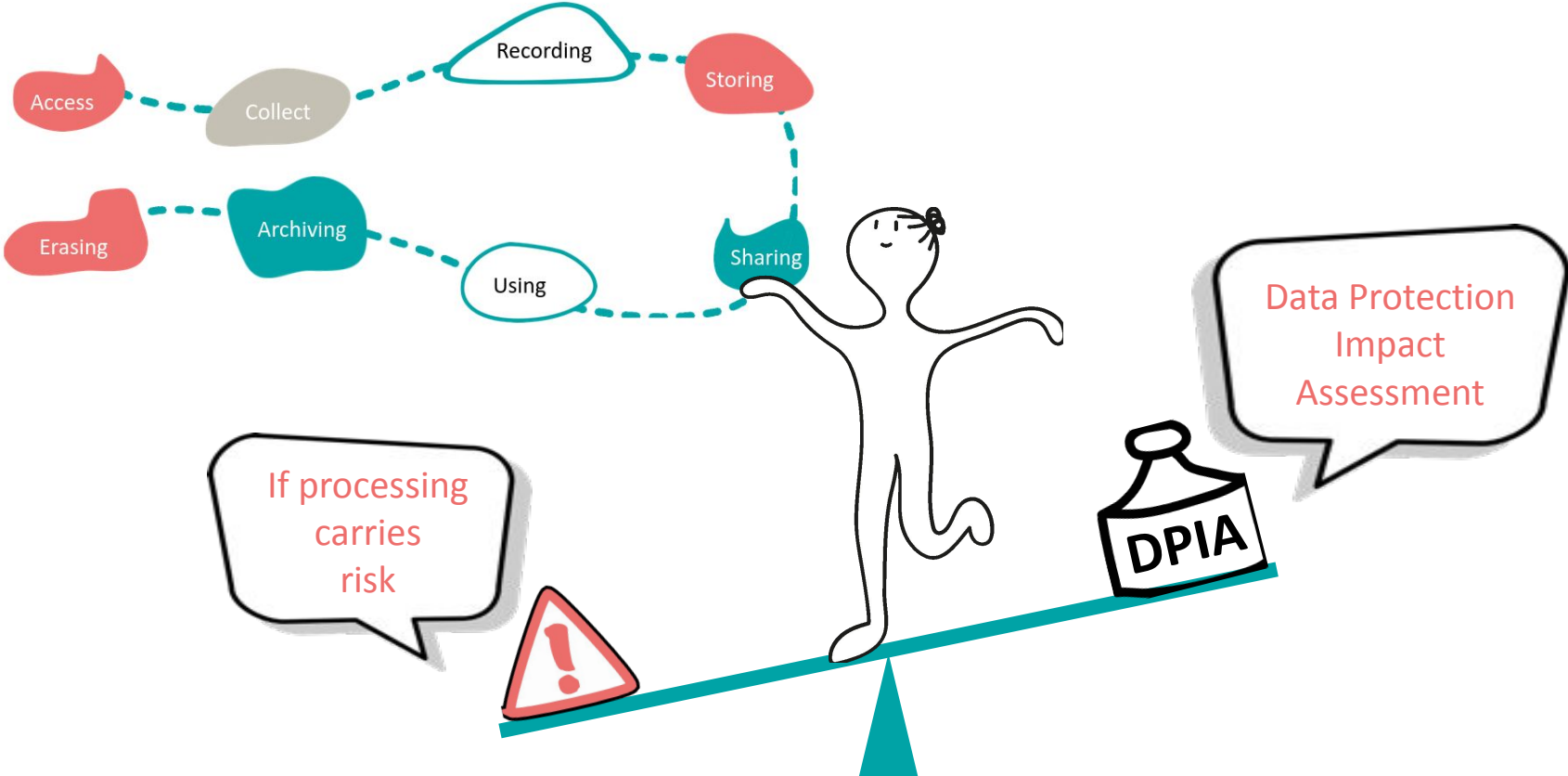
GDPR Main Concepts: more responsibilities

Accountability principle: Change in practices, not in principles



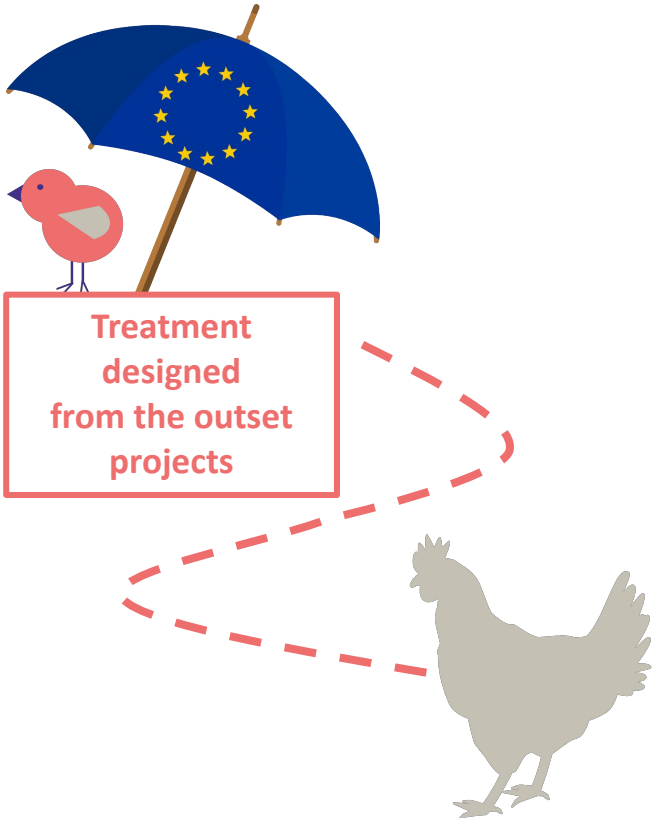
GDPR Main Concepts: more responsibilities

Privacy Impact Assessment



GDPR Main Concepts: more responsibilities

Privacy by design



5 principles for the respect privacy

Specific and legal purpose



Minimising collect



Security



Limited data retention periods

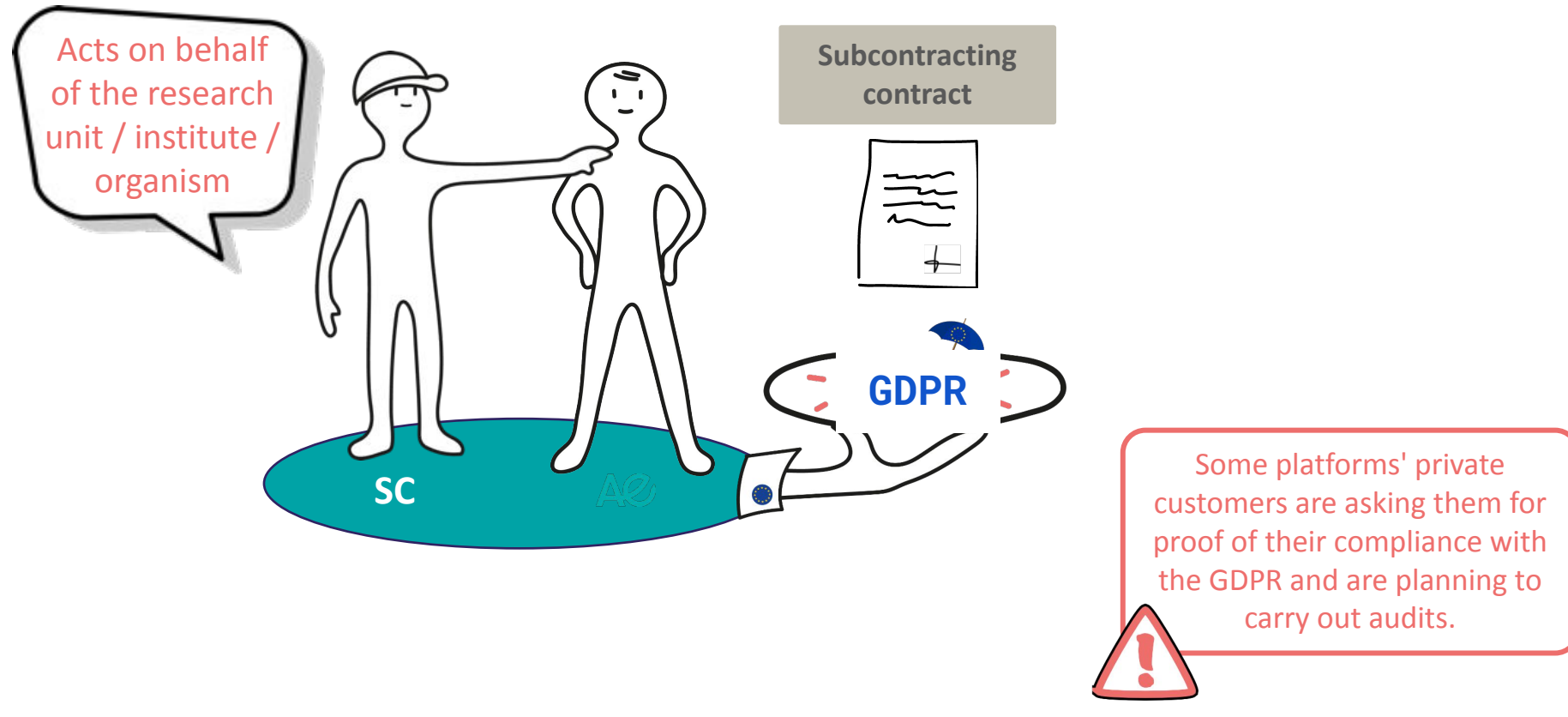


Individual rights



GDPR Main Concepts: more responsibilities

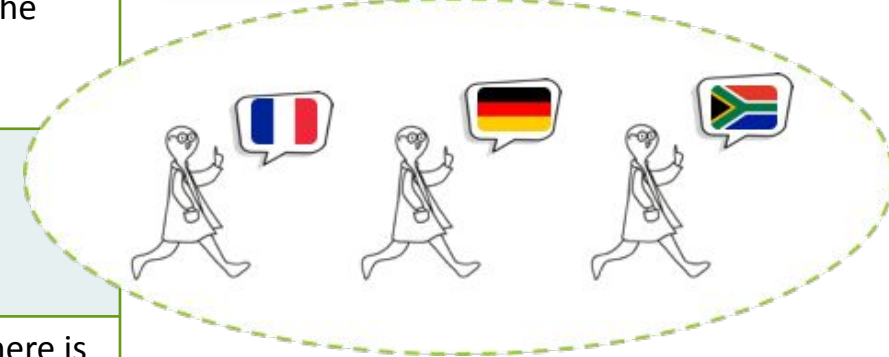
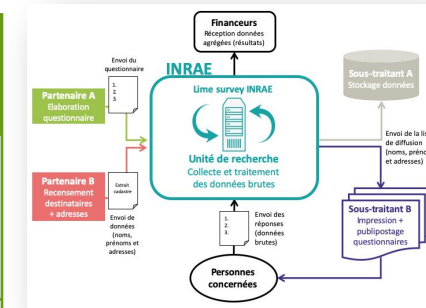
New status for subcontractors



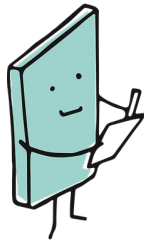
GDPR - Research project - Questions to ask




Subject		
1	Role of partners	Who handles PD and how? Diagram the flow of data between partners
	Check partners' nationality	If appropriate: OK If not: contract complementary to the consortium agreement
	Joint liability agreement between partners	Even between European partners
	Verification of the place of residence of the people whose data you are going to process	If outside Europe, check whether there is a national law on personal data: to be checked with local partners



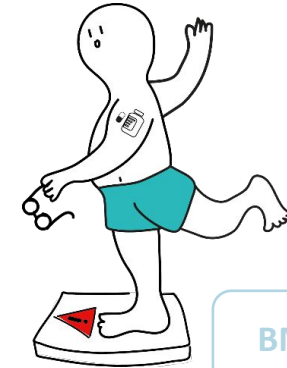
GDPR - Research project - Questions to ask



Subject	
2	<p>Legal basis for processing PD</p> <p>Depending on the purpose and partners (public, private): consent, public interest mission, legitimate interest</p>
3	<p>Ethics committee</p>  <p>Certain types of project (particularly research involving humans) may need to be submitted to an ethics committee. Some journals also require a favourable opinion from the committee before publishing an article.</p>
4	<p>People selection</p> <p>Transparent inclusion/exclusion criteria</p>
5	<p>Nature of PD handled and level of impact (1 / no risk 4 / high risk)</p> <p>Possible justification? Remember to use "categories" for answers</p>
6	<p>Reuse of pre-existing PD</p> <p>From which partner? Will you need to <u>re-inform people</u> or not?</p>

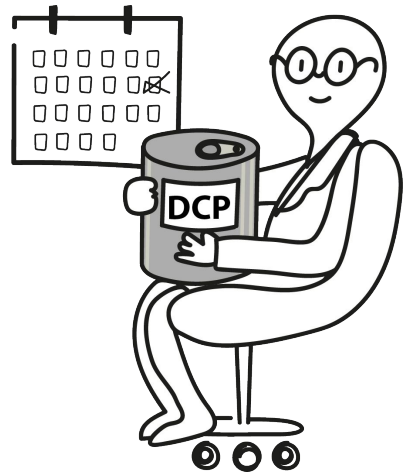


Please note!
Different rights depending on the legal basis chosen

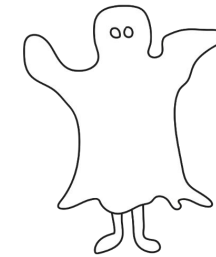


BMI > 30 

GDPR - Research project - Questions to ask



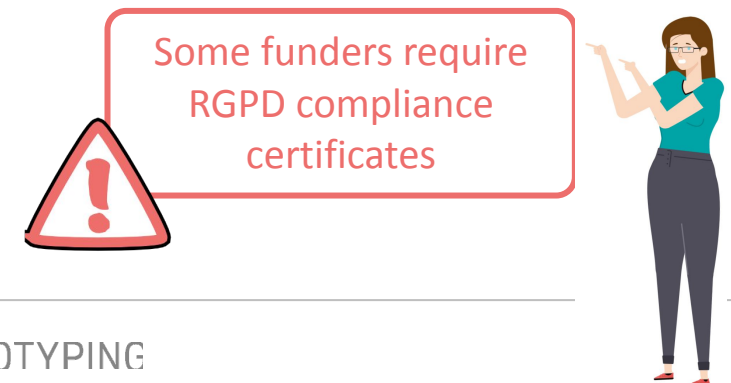
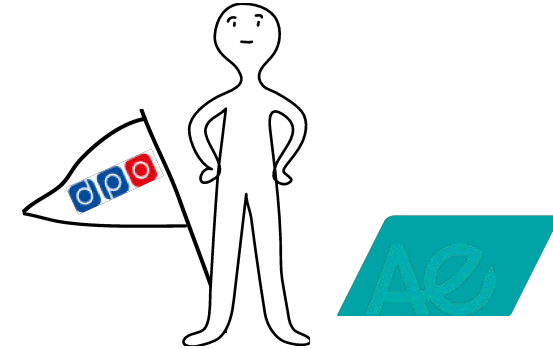
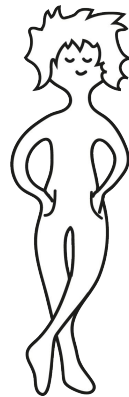
Subject	Questions et actions
7 Tools used	<p>What collection tools? For storage? Processing?</p> <p>Is it subcontracted? Read the GCU/GTC</p> <p>Does it involve a transfer outside the EU (subcontractor's parent company outside the EU or appropriate country)?</p>
8 IT security	<p>To be adapted according to the nature of the data. If level 3 to 4: risk analysis</p> <p>Storage, access procedure</p>
9 Retention period and future of data	<p>What is the useful life until future publications?</p> <p>What happens to the data (anonymisation / archiving / destruction)?</p>
10 Information and people's rights	<p>If consent: proof and withdrawal procedure to be provided for</p> <p>Generic alias for exercise of rights</p>



GDPR - Research project - Questions to ask



Subject	Questions et actions
Registering the project	Internal process for each partner
Compliance	Variable duration - 1 month for simple cases <input type="checkbox"/> Take into account the complexity of the project and the nature of the risks to people.



Some funders require RGD compliance certificates

GDPR - Research Project Good Practice



Sujet	Questions et actions
Open data	<p>PD may not be made available in open data unless :</p> <ul style="list-style-type: none"> - the persons concerned have given their consent, - a legislative or regulatory text allows it (such as a decree on the publication of certain administrative documents without anonymisation), - the data is rendered anonymous, so it is no longer personal, but there will be the inevitable loss of all the criteria enabling individuals to be identified, particularly by cross-referencing.
The project has a website	Don't forget the legal notices, terms and conditions and information on cookies.
Non-permanent staff	Get them to sign a <u>confidentiality agreement</u>
Transcription / translation of interviews	In addition to the order form for the transcription/translation service, it is preferable to have an GDPR rider signed.



Non-compliance GDPR: What are the risks?

Examples of offences:

- Treatment despite opposition of the individual
- Diversion of purpose
- Lack of impact assessment
- Failure to cooperate with the supervisory

Image impact

Public Warning
Loss of confidence as a partner establishment
Loss of appeal to volunteers

Financial impact

Criminal and administrative fines
Compensation for damages

Conformity impact

Stop treatment (=stopping research)
Criminal conviction



Legal action possible via group actions



Research project - Questions to ask...

... not only for personal data!!

GDPR requires us to ask ourselves questions about personal data...

... but these questions are important to ask for all project data!!

Conclusion

FAIR Data, DMP, GDPR...

- If you are thinking about FAIR data, you are building your DMP
- If you're thinking about your DMP, you're in the process of setting up FAIR management.
- In any case, the DMP is a mandatory part of all projects, so you might as well do it properly.
- The GDPR is a legal formalisation designed to protect the rights of individuals, but most of the questions and points of attention raised are of interest for all types of data.

Thank you for your attention!



<https://www.phenome-emphasis.fr/>



OpenSILEX Team - <http://opensilex.org/>

Special thanks to: Silvana Moscatelli, François Tardieu, Renaud Colin, Célia Michotey, Frédéric De-Lamotte, Cyril Pommier, Nathalie Gandon, Adrien Mousset, ...

 emphasis@fz-juelich.de

 emphasis.plant-phenotyping.eu

 EMPHASIS_EU

 EMPHASIS.EU

 EMPHASIS on Plant Phenomics



EMPHASIS is an ESFRI-listed project.



EMPHASIS-PREP is funded by the European Union (Grant Agreement: 739514).

EUROPEAN INFRASTRUCTURE
FOR PLANT PHENOTYPING