Day 6: Data management and ethics

Guiding Principles for Data Management I

Sebastian Ramirez-Ruiz
Hertie School

What people think working with data looks like...

Hertie School

How it really is...

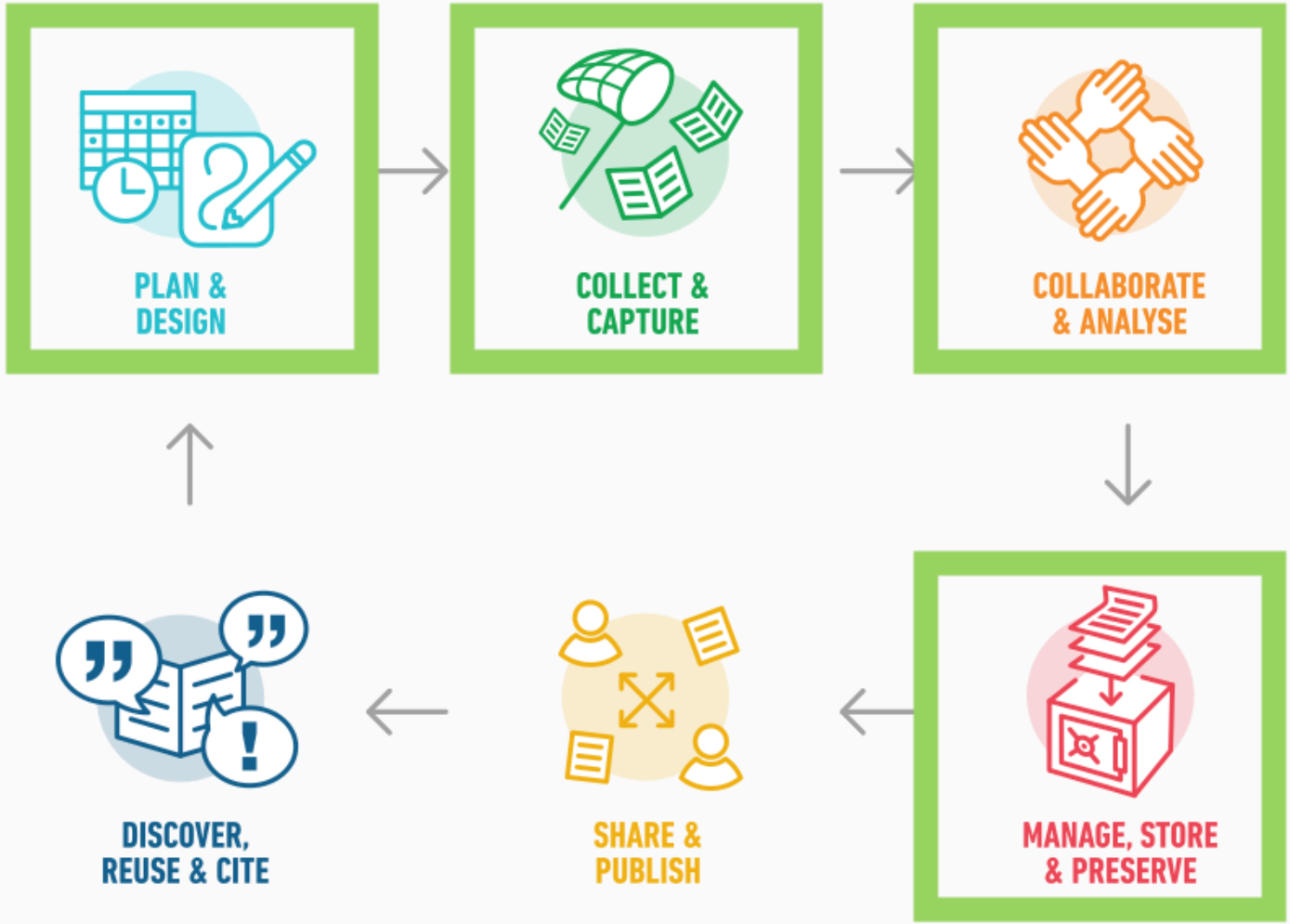
You have to wear many hats...

Research data management (RDM) lifecycle

Hertie School

Table of contents

Research data management (RDM) lifecycle



A show of hands!

Raise your hand if you:

- have heard the term data management plan (DMP)
- can define what a data management plan entails
- have ever created a data management plan for a project
- have planned data collection methods for research or policy work
- have implemented strategies to ensure data integrity and security
- are familiar with the FAIR principles (Findability, Accessibility, Interoperability, and Reuse)
- have incorporated open data practices into your work
- have archived or published datasets for public access
- are familiar with data protection regulations and compliance requirements
- have used techniques to discover and access relevant datasets
- have utilized data management tools or software in your projects
- have trained or guided others in effective data management practices

Data management and a data management plan (DMP)

What are data?

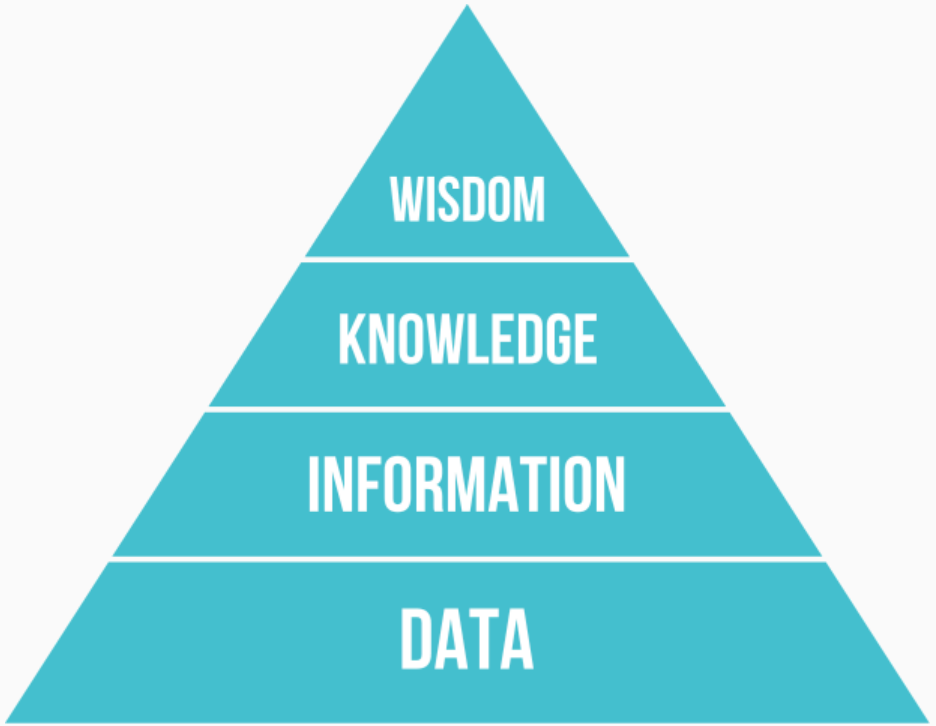
- Data are disembodied **facts**, **signs**, and **symbols**.

What are data?

- Data are disembodied **facts**, **signs**, and **symbols**.
- We often define them by their **source** (*e.g. administrative, historical, medical, etc.*) and their **formats** (*e.g., numerical, textual, still image, geospatial, audio, video, and software.*)

What are data?

- Data are disembodied **facts**, **signs**, and **symbols**.
- We often define them by their **source** (*e.g. administrative, historical, medical, etc.*) and their **formats** (*e.g., numerical, textual, still image, geospatial, audio, video, and software.*)
- It is thought about as the basis of the **knowledge hierarchy** under the *DIKW pyramid*.



Data in policy analysis

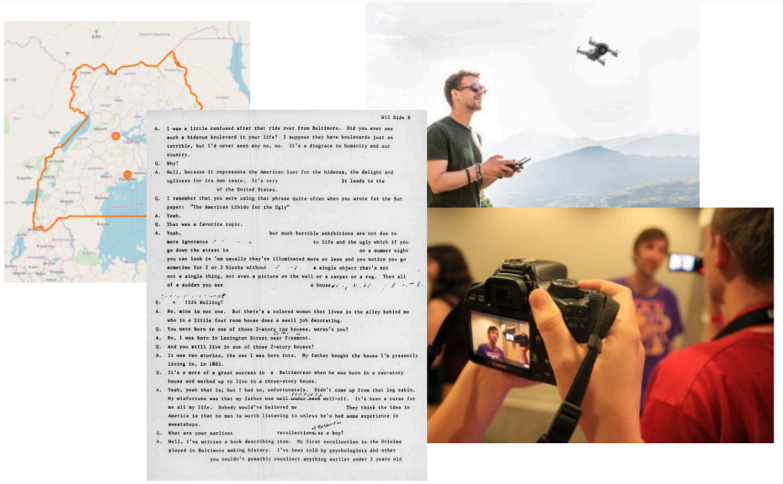
In a similar way to academic research, as **policy analysts you might rely on a broad range of materials**, from structured numerical datasets to interviews, field notes, and documents collected for ethnographic field studies,

Quantitative

startdate	startdate2	startdate3	lastupdate	lastupdate2	lastupdate3	collectorcreator
13599588513	26-Sep-2013 15:28:33	2013-09-26 15:29:01.440	13599589090	26-Sep-2013 15:38:10	2013-09-26 15:37:45.728	DIGITA08
13598878173	18-Sep-2013 10:09:33	2013-09-18 10:08:51.200	13598878822	18-Sep-2013 10:20:22	2013-09-18 10:19:46.560	DIGITA07
13598879269	18-Sep-2013 10:27:49	2013-09-18 10:28:30.848	13598879836	18-Sep-2013 10:37:16	2013-09-18 10:37:15.136	DIGITA07
13598879940	18-Sep-2013 10:39:00	2013-09-18 10:39:26.208	13598880525	18-Sep-2013 10:48:45	2013-09-18 10:48:10.496	DIGITA07
	18-Sep-			18-Sep-		

Viewing rows 6 through 9 of 1511

Qualitative



Regardless of *"type"*, in the policy world, we will likely deal largely with **collection of data about individuals** (i.e., *human subject research*).

Personal data

Any data that enables you to **identify a person** is classified as personal data. In the General Data Protection Regulation (GDPR) personal data are *any information relating to* an **identified** or **identifiable** natural person known as 'a data subject'[1].

Sensitive personal data

Certain personal data can *require specific protection* when they reveal information that may create important *risks* for the *fundamental rights* and *freedoms* of the involved individual. In the context of GDPR, these can be:

- Racial or ethnic origin;
- Political opinions;
- Religious or philosophical beliefs;
- Trade union membership;
- Genetic data;
- Biometric data;
- Data concerning health;
- Data concerning a natural person's sex life or sexual orientation.

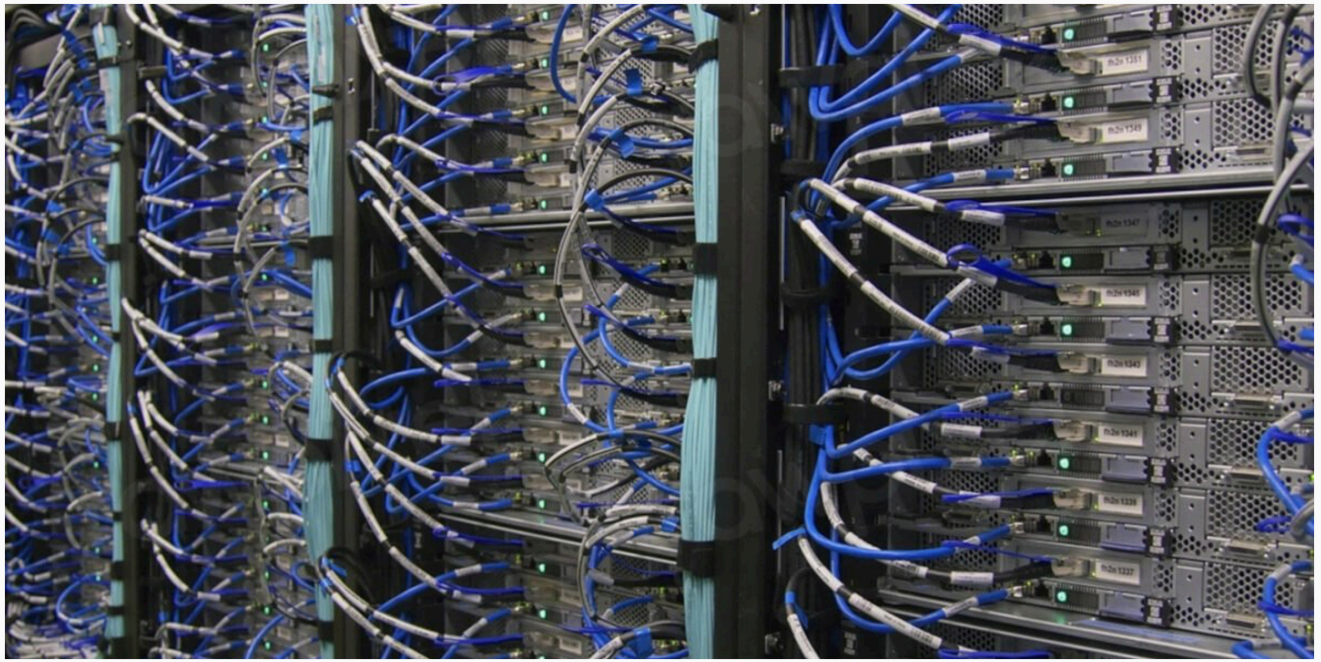
[1]The GDPR applies only to the data of living persons. Data which do not count as personal data do not fall under data protection legislation, though there might still be ethical reasons for protecting this information.

Research data management (RDM)

Research data management is like "*health care*" for your data:

- keeps them safe from harm,
- makes them usable and discoverable.

It entails strategies, processes and measures to maintain data quality, interpretability of research results, (re-)usability of your data.



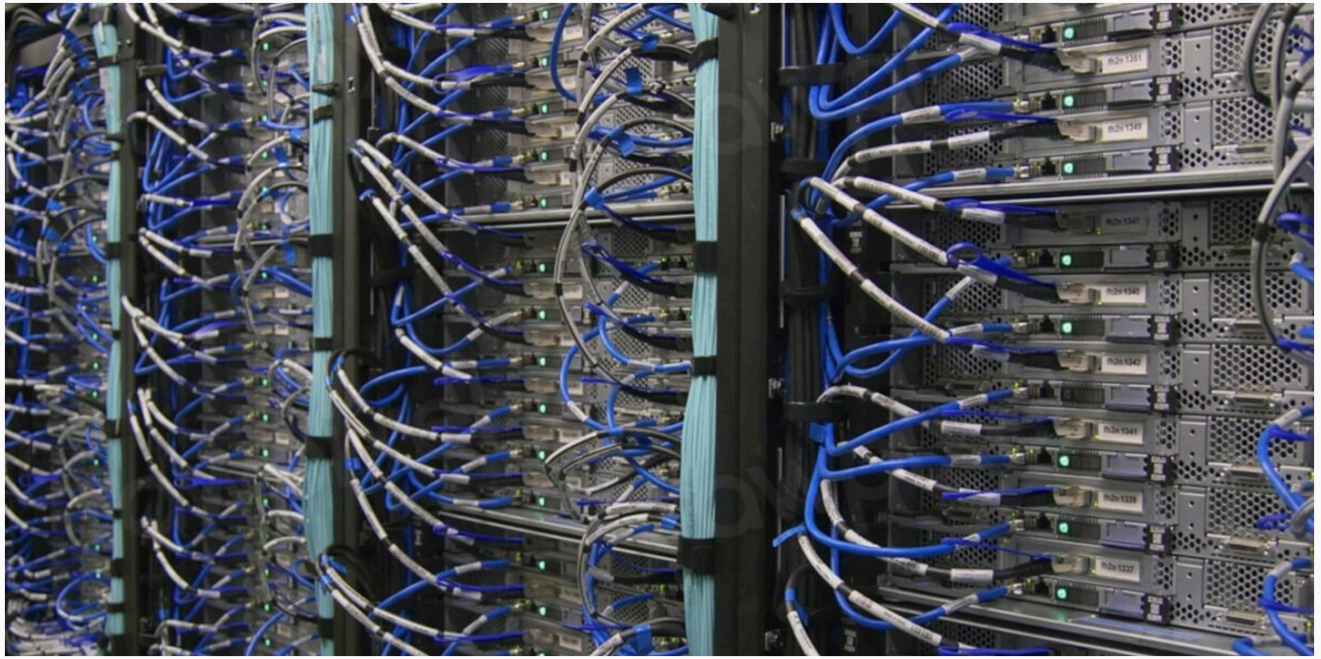
Research data management (RDM)

The *need* for thinking closely about **RDM** increases when dealing with *sensitive personal data*.

Research data management is like "*health care*" for your data:

- keeps them safe from harm,
- makes them usable and discoverable.

It entails strategies, processes and measures to maintain data quality, interpretability of research results, (re-)usability of your data.



Research data management (RDM) lifecycle

Data management plan (DMP)

Data management plans are an important tool to structure the data management of your project.

Defines strategies, measures and responsibilities for:

- processing and validating,
- storing and protecting,
- preserving and sharing

your data throughout the data cycle.

More and more *funding partners increasingly require them*.

Components of a data management plan

1. Data summary
2. FAIR data
    - *Findable*
    - *Accessible*
    - *Interoperable*
    - *Reusable*
3. Allocation of resources
4. Data security
5. Ethical aspects
6. *Other**

Data management plan (DMP) (cont.)

DMP component	Issues to be addressed
1. Data summary	- State the purpose of the data collection/generation - Explain the relation to the objectives of the project - Specify the types and formats of data generated/collected - Specify if existing data is being re-used (if any) - Specify the origin of the data - State the expected size of the data (if known) - Outline the data utility: to whom will it be useful
FAIR data: **2.1. Making data findable, including provisions for metadata**	- Outline the discoverability of data (metadata provision) - Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? - Outline naming conventions used - Outline the approach towards search keywords - Outline the approach for clear versioning - Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how

`Source`: H2020 DMP Template

Data management plan (DMP) (cont.)

DMP component	Issues to be addressed
FAIR data: **2.2 Making data openly accessible**	- Specify which data will be made openly available? If some data is kept closed provide rationale for doing so - Specify how the data will be made available - Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? - Specify where the data and associated metadata, documentation and code are deposited - Specify how access will be provided in case there are any restrictions
FAIR data: **2.3. Making data interoperable**	- Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. - Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

`Source`: H2020 DMP Template

Data management plan (DMP) (cont.)

DMP component	Issues to be addressed
FAIR data: **2.4. Increase data re-use (through clarifying licences)**	- Specify how the data will be licensed to permit the widest reuse possible - Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed - Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why - Describe data quality assurance processes - Specify the length of time for which the data will remain re-usable
3. Allocation of resources	- Estimate the costs for making your data FAIR. Describe how you intend to cover these costs - Clearly identify responsibilities for data management in your project - Describe costs and potential value of long term preservation
4. Data security	- Address data recovery as well as secure storage and transfer of sensitive data
5. Ethical aspects	- To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former
6. Other	- Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

`Source`: H2020 DMP Template

Protecting

Research ethics

The moral **principles** and **actions** guiding and shaping research

In policy analysis is closely linked to research ethics in the social sciences.

- Initially, a 'patient protection' model from medical research.
- Today it has a broader scope, including:
  - consideration of *benefits, risks, and harms*
  - *to all persons* connected with and affected by the research
- Gives **responsibilities** to researchers and analysts (e.g., *legal framework*).

Nuremberg code (1947)

Ethical guidelines for the preparation and conduct of medical, psychological and other experiments on humans:

> "The voluntary consent of the human subject is absolutely essential [...] without the intervention of any element of force, fraud, deceit, duress, over-reaching, or other ulterior form of constraint or coercion [...] should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an **understanding and enlightened decision**."

Reviews of ethics in academic research

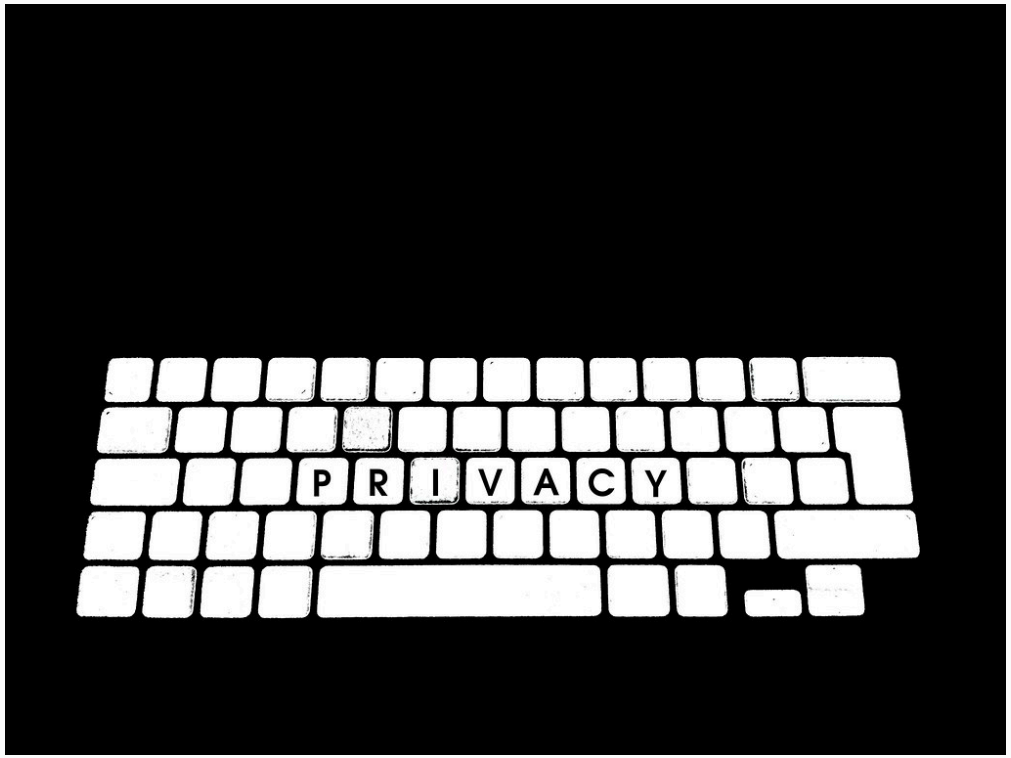
Institutional Review Board (IRB) and ethical boards

- Administrative bodies established to **protect the rights and welfare** of research subjects
- Often housed within the academic institution
- Checks proposals and recommends changes and improvements
- *Can kill unethical projects by denying approval*

What is "data protection"?

Data protection

- Part of fundamental *right to privacy* (or 'informational freedom').
- In research there might be instances of tensions of fundamental rights:
  - Freedom of research vs. freedom of personal information.
- “Privacy is a personal condition of life characterized by seclusion from, and therefore absence of acquaintance by, the public” (Neethling 2005).
- **Core**:
  - prevention of unwanted disclosure of personal information or
  - the misuse of such information.



There are individuals behind the data...

Data protection principles under GDPR

Article GDPR	Topic	Meaning
Art 5 (1) (a)	Lawfulness	Data must be processed in a legal way (Art 6) and transparent for 'data subjects'; no surprises or covert activities.
	Fairness	
	Transparency	
Art 5 (1) (b)	Purpose limitation	Data may only be collected "for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes"; research exemption: research seen as in line with initial purposes.
Art 5 (1) (c)	Data minimisation	Limit amount of data collected.

Data protection principles under GDPR (cont.)

Art GDPR	Topic	Meaning
Art 5 (1) (d)	Accuracy	Data collected for a given purpose should be kept correct and deleted or corrected without delay if necessary.
Art 5 (1) (e)	Storage limitation	Research exemption: longer period, if “appropriate technical and organizational measures” are implemented.
Art 5 (1) (f)	Integrity	Protected against “unauthorized or unlawful processing and against accidental loss, destruction or damage”.
	Confidentiality	
Art 5 (2)	Accountability	Controller (or processor) in charge and liable.

General Data Protection Regulation

- In place since May 25, 2018 (*almost six years to the day* 😎)
    - 99 articles and 173 recitals
    - Applies directly
    - Intended to harmonize data protection law EU-wide
    - *but*, about 150 “opening clauses” or exemptions...
- GDPR (factually) integrated into a hierarchy of norms

[1] Let's check this piece out https://www.bbc.com/news/uk-wales-politics-58395974

Organizing, documenting, processing, and storing

Data processing under GDPR

Article 6 (1) GDPR	Examples
a) Consent	
b) Performance of a contract	Employment contracts; "two-sided obligational relationship" (e.g., sales contract); membership in association; *contract with the person concerned!*
c) Compliance with a legal obligation	Legal obligation to process data; e.g. documentation obligations under commercial law or notification obligations under social security law.

Data processing under GDPR (cont.)

Article 6 I GDPR	Examples
d) Protection of vital interests	*Immediate threat* to vital interests of people exists; e.g., humanitarian emergencies and catastrophes.
e) Public interest or exercise of official authority	Tasks in the public interest; e.g., health, or justice and law enforcement.
f) Legitimate interests	Most important legal basis; advantage > flexibility; higher risks of data processing, because sole base on the initiative of the controller; e.g., data in companies, internet, credit agencies.

Informed consent for data processing

Art 4 Par 11 GDPR

> "'(C)onsent' of the data subject means **any freely given, specific, informed and unambiguous indication of the data subject's wishes** by which he or she, by a statement or by a clear affirmative action, signifies **agreement to the processing of personal data** relating to him or her;"

Conditions for consent

- Written form no longer required.
- Conditions for consent (Art 7 GDPR):
  - Requirement to provide evidence (Par1)
  - Separation requirement (Par2)
  - Easy revocability at any time (Par3)
  - Increased requirement of voluntariness (Par4)
- More stringent for minors under 14 years of age (Art 8 GDPR).
- Important: Collection of "special categories" of personal data "prohibited" under Art 9 GDPR, unless there is a legal basis.

- **Central**: *consent needs to be documented!*

Can we process data without consent? (in research)

Are researchers allowed to process data that was collected without consent?

- *It depends...*
- May need to try to gain consent after the data was gathered.
- They may skip this step, if "the provision of such information proves impossible or would involve a disproportionate effort" (Art 14 Par 5 Lit b).
- But this is no default!
- **Disproportionate really means disproportionate**.

Consent after the fact is a balance of interests

Consent after the dact and the balance of interests

Hertie School

"Balance of interest" - checklist:

- Legitimate interest of the researcher?
- Is the data processing necessary?
  - Are the more lenient ways to achieve research goal?
- Can the subject refuse to the processing?
- Are these data linked (or linkable) to other data?
- How long will the data be stored?
- How many people are going to access the data?
- Are the research subjects from a vulnerable group?
- Would the research subjects have to expect the processing of their data?

THINK CLOSELY: if research subjects' interest weigh more than yours, data processing is improper

Secure storage

Technical and organizational measures (TOMs)

- **The security of out data**

> TOMs should be designed to "implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects" (Art 25 Par 1 GDPR).

Examples of TOMs

Measure	Implementation
Data avoidance	Only collect as much data as necessary.
Data separation	Keep e.g. survey data (audio files, transcripts, ...) and contact data separate.
Admission control to building / office	E.g. lock office after leaving.
Access control to computer	E.g. protect PC with personal password.
Access control to files	E.g. agree on access to the electronic filing system and keep it as low as possible.

Pseudonymization and anonymization

Pseudonymization

(Art 4 Par 1 No 5 GDPR):

> "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;"

With pseudonymization the possibility of establishing the true identity remains

Anonymization

(GDPR Recital 26):

> "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

With anonymization there is no possibility of establishing the true identity

What is enough?

- This is a question that **does not have a clear answer**.
- Consider the **risk-based approach** of GDPR.
- Technical and organizational measures (TOMs) aid you in keeping the risk at bay.
- If working with large datasets, consider means of **Statistical Disclosure Control** (e.g. k-anonymity).
- If you are uncertain, turn out to a **data management and legal expert**!!!

Data security

Does anyone know the difference between *data protection* and *data security*?

Data security is more encompassing and not necessarily person related.

- Encryption
- Strong passwords
- Access rights
- Back ups

Encryption

Encryption maintains security of data

- Uses an algorithm to transform information
- Needs a "key" to decrypt

Use encryption to

- Transfer data
- Store data (back-ups)
- On remote discs

Tools. e.g., 7Zip, Gpg4win, Veracrypt

Here is some extra information about encryption by the Federal Office for Information Security 🇩🇪

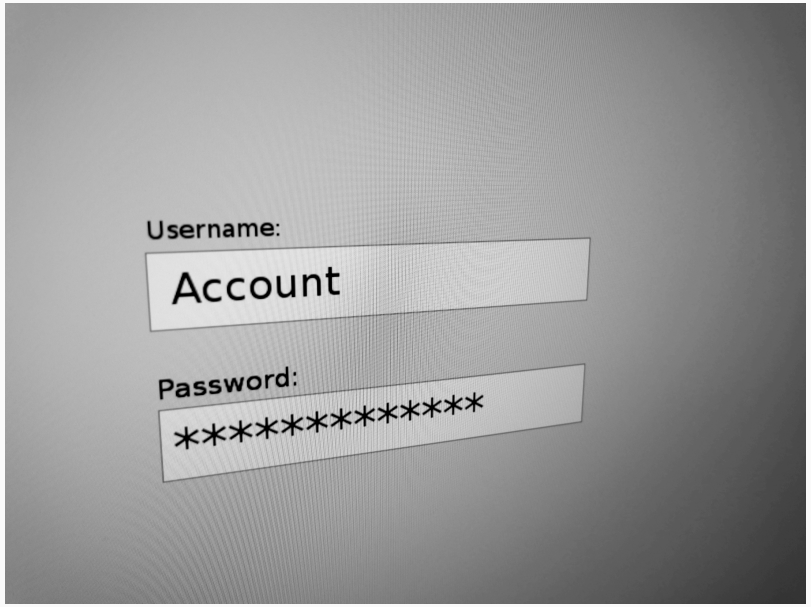
Strong passwords

A strong password has:

- eight to fifteen characters or even more
- a random distribution of characters

Combine

- upper case letters: A - Z
- lower case letters: a - z
- numerals: 0-9
- special characters: !"#$%&'()*+,-./:, etc.



Use a ‘pass-sentence’ instead of a password!

Here is some extra information about password managers by the Federal Office for Information Security 🇩🇪

Secure back-ups

Risks:

- Technical defects
- Catastrophes
- Theft
- Forgetfulness

Strategies

- Storage on secure servers with automatic regular backup
- Backup important files in at least three copies on spatially separated data carriers

Backup setup (3-2-1 rule)

- At least **3** copies of a file
- On at least **2** different media
- At least **1** of which is remote

Test data recovery at the beginning and at regular intervals

Protect your (sensitive) data:

- Hardware (e.g. separate lockable room)
- File encryption
- Password security
- At least two people should have access to your data

Here is some extra information about data backup by the Federal Office for Information Security 🇩🇪

Why this matters to us?

Let's take a look at a couple of cases

- **What happened?**
- **Why was it a problem?**
- **What should have happened?**

UK Information Commissioner's Office (ICO)

Questions?