FORS Data and services for the social sciences

DaSCH

Data Management Planning

Workflows for Anonymization

FORS and DaSCH

FORS – Swiss Centre of Expertise in the Social Sciences

- national infrastructure for Social Sciences' research data mainly funded by SNSF
- services: consulting / training / workshops /events for data management and archiving, SWISSUbase repository for the social sciences, mandates around the data collection and analysis, FORS Guides

DaSCH – Swiss National Data and Service Center for the Humanities

- national infrastructure for Humanities' research data mainly funded by SNSF
- services: consulting / training / workshops /events, virtual research environment, FAIR open data repository (DSP) including data publication and persistent identifiers at object level, metadata browser

Rita Gautschy (DaSCH)
rita.gautschy@dasch.swiss

Pedro Araujo (FORS)
pedro.araujo@unil.ch

Programme

01 Conceptual and legal foundations of anonymisation

02 Anonymisation in practice: trade-offs and limitations

03 Humanities

04 Images: between documentation and exposure

05 Ethnology & Indigenous data

06 Useful tools for anonymisation and pseudonymisation

07 Q&A

Conceptual and legal foundations of anonymisation

What is anonymisation?

- **Definition**
  *Anonymisation refers to a process by which the elements allowing the identification of a person are **definitively** deleted or modified, making identification impossible, or at least very difficult*
- Anonymised data ≠ **personal data**
- Anonymisation is carried out to **enable publication** and **data sharing**, within a **legal** and **ethical framework**

Anonymisation is context-dependent

- Every research project has its own data context. Especially in SSH, what counts as “**data**” varies widely.
- Key questions before starting anonymisation:
  - What type of data do I have? (survey, interviews, observations, audio, video, images, …)
  - Have you established an informed consent form? What commitments did you make?
  - What identifiers are present in this data? (direct and indirect identifiers)
  - What are the legal and ethical risks associated with data publication and sharing?
  - What can be shared and under which constraint(s)?

What counts as personal data?

- What qualifies as personal data is determined by applicable **data protection law**
- In the Swiss research context, two legal frameworks are particularly relevant:
    - The EU General Data Protection Regulation (**GDPR**)
    - The Swiss Federal Act on Data Protection (**nFADP**)
- Under these laws, personal data means:
  “***All information*** *relating to an* ***identified or identifiable individual***” (Art. 5 let.a nFADP)

Interpreting “personal data”

- **“All information”**
  - The notion must be interpreted as broadly as possible by researchers (*e.g.* text, voice, image)
- **“Identified or identifiable individual”**
  - Information can identify a person **alone or in combination** with other data
  - A person my be “identified” **directly** or “identifiable” **indirectly**

Direct identifiers & indirect identifiers

- **Direct identifiers**: information that identifies an individual directly
  - Examples: name, phone number, image, voice, address, ID number
- **Indirect identifiers**: combinations of attributes that may identify an individual when combined
  - Examples: (1) profession + location + age (2) ethnicity + sex + small population context
- Anonymisation, therefore consists in addressing both direct and indirect identifiers in order to **reduce the risk of identification**

Identifiability and levels of risk

- The level of identification risk varies depending on the nature of the data
- Risk assessment involves both the likelihood of identification and the severity of potential consequences
- Example of **low-risk identification**
    - A poorly anonymised survey on eating habits in a university cafeteria reveals that a professor prefers chocolate desserts
- Example of **high-risk identification**
    - A sociologist interviews undocumented migrants in a small municipality. Identification may expose the participants to legal and life threatening consequences
- Some categories of personal data are legally considered more sensitive and require enhanced protection

Personal data vs. sensitive data

- Data protection law distinguishes between **personal data** and **sensitive (personal) data**
- Under the Swiss nFADP (art. 5 let. c), sensitive data are personal data relating to:
    - Religious, philosophical, political or trade union-related views or activities
    - Health, the private sphere, racial or ethnic origin
    - Genetic data
    - Biometric data
    - Administrative and criminal proceedings or sanctions
    - Social assistance measures
- The more sensitive the data, the stricter the legal and ethical requirements. In SSH research, depending on the context, almost all personal data can be considered sensitive.

FORS Data and services for the social sciences

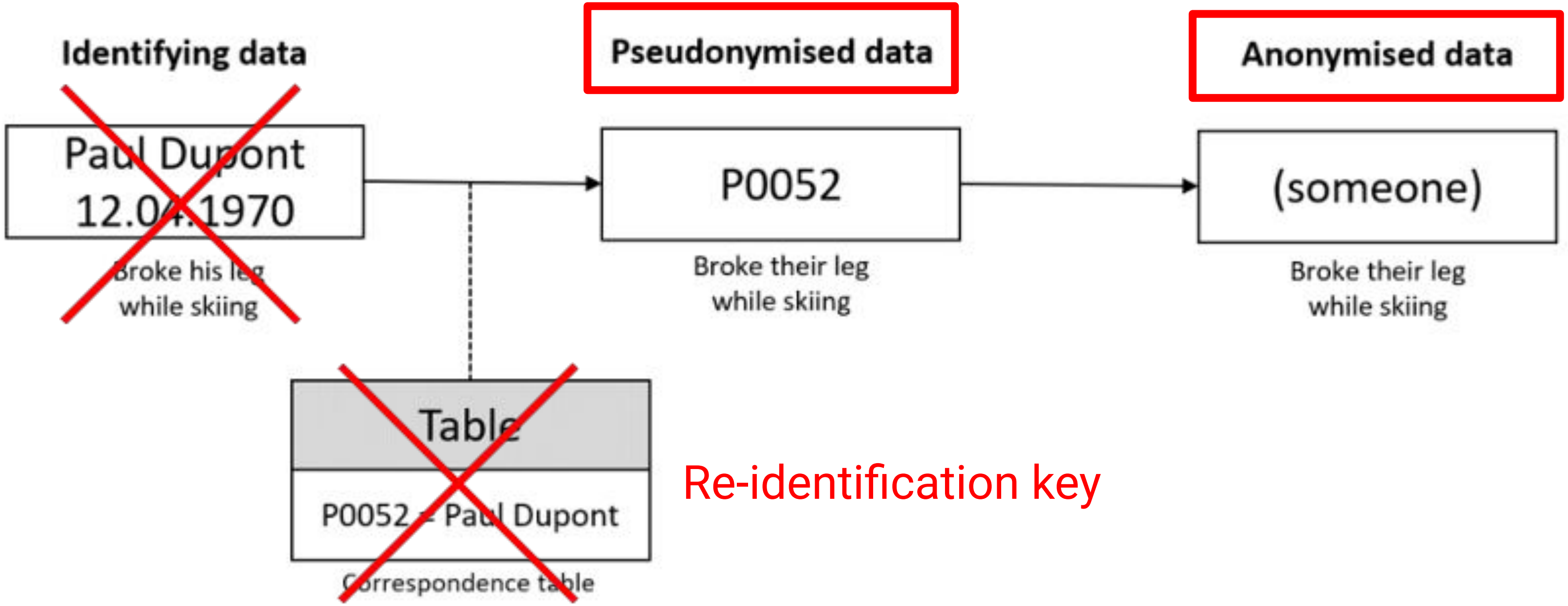
DaSCH

Anonymisation in practice: trade-offs and limitations

Anonymisation vs. Pseudonymisation

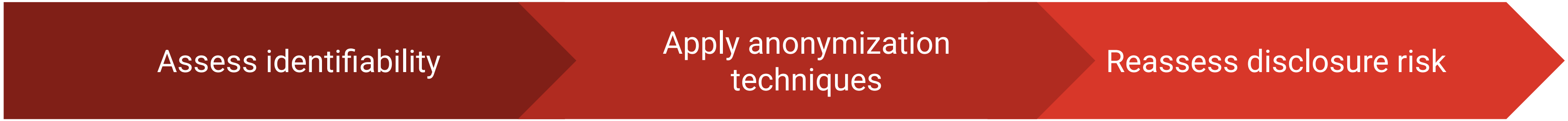
- **Anonymisation**: data are processed so that individuals are no longer identifiable by any means (irreversible)
  - Data are no longer "personal data" and fall outside data protection law
- **Pseudonymisation**: identifiers are removed, masked, replaced by codes or pseudonyms, but re-identification remains possible via a **key**
- In SSH, many datasets described as "anonymous" are in fact pseudonymised

Anonymisation vs. Pseudonymisation



Source: RACINE, Céline (Unisanté) & STAM Alexandra (FORS), 2024. Legal Bases – Anonymisation workshop [Lausanne]. CAS Data Stewardship UNIL. 31 October 2024.

Anonymisation as a process



Assess identifiability

Apply anonymization techniques

Reassess disclosure risk

- Identify direct and indirect identifiers
- Consider context and population size
- Evaluate the disclosure risk

- Remove direct identifiers
- Transform indirect identifiers
- Document decisions

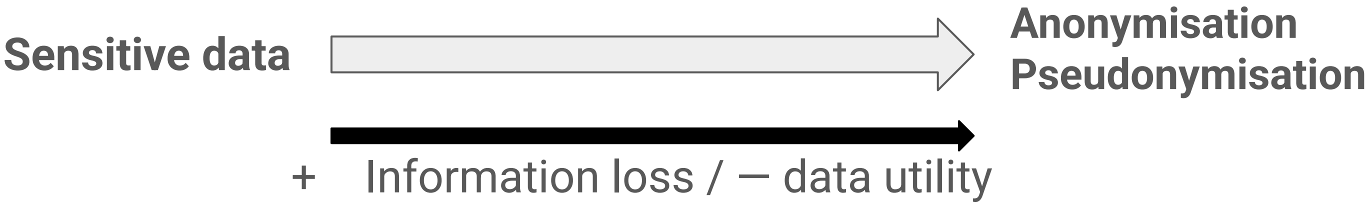
- Assess if identification is still reasonably possible
- Assess potential harms
- Assess control access need (always recommended for qualitative data)

Common anonymisation techniques (quick overview)

- **Suppression**: deleting identifiers, in priority direct identifiers (names, IDs, face, …), sensitive open-ended questions, comments, unuseful information
- **Pseudonymisation techniques:** replacing identifiers with artificial values
  - Replace names with codes or fictitious names
- **Generalisation**: reducing the precision of data
  - **Recoding** variables to reduce categories, transform continuous variables into discrete ones
  - **Categorisation** into broader categories (*e.g.* age *36* -> age range *35-39*; municipality-> region)

The anonymisation trade-off

- More anonymisation = lower identification risk
- But more anonymisation = less data utility



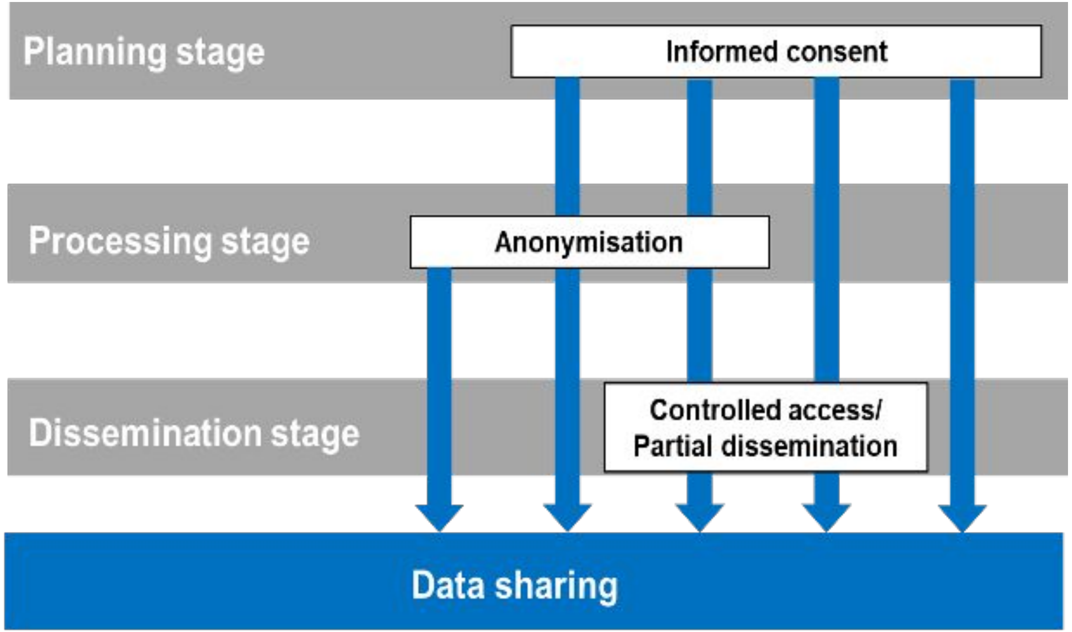
Sensitive data

Anonymisation
Pseudonymisation

+ Information loss / — data utility

- Anonymisation is **not neutral** it reshaped empirical material
- Effective data sharing requires balancing: anonymisation with **informed consent** and **controlled access**

A layered approach to data protection



Humanities

Why anonymisation matters in the humanities

Humanities data can be

- deeply personal
- politically sensitive
- culturally restricted
- social consequential

Not necessarily “personal data” in a narrow legal sense!

Anonymisation in the humanities is not merely a technical or legal task, but an ethical, contextual, and relational practice.

Why anonymisation matters in the humanities

Anonymisation vs. Pseudonymisation

- Full anonymisation is often impossible or undesirable
  - Data would lose its value (e.g. context, authority, historic specificity lost)
- Pseudonymisation, controlled access, and contextual restriction often more appropriate strategies

Be aware:

May not be a one-time-action: data that appears anonymous today may become identifiable tomorrow through technological advances.

Images
Between documentation and exposure

Images

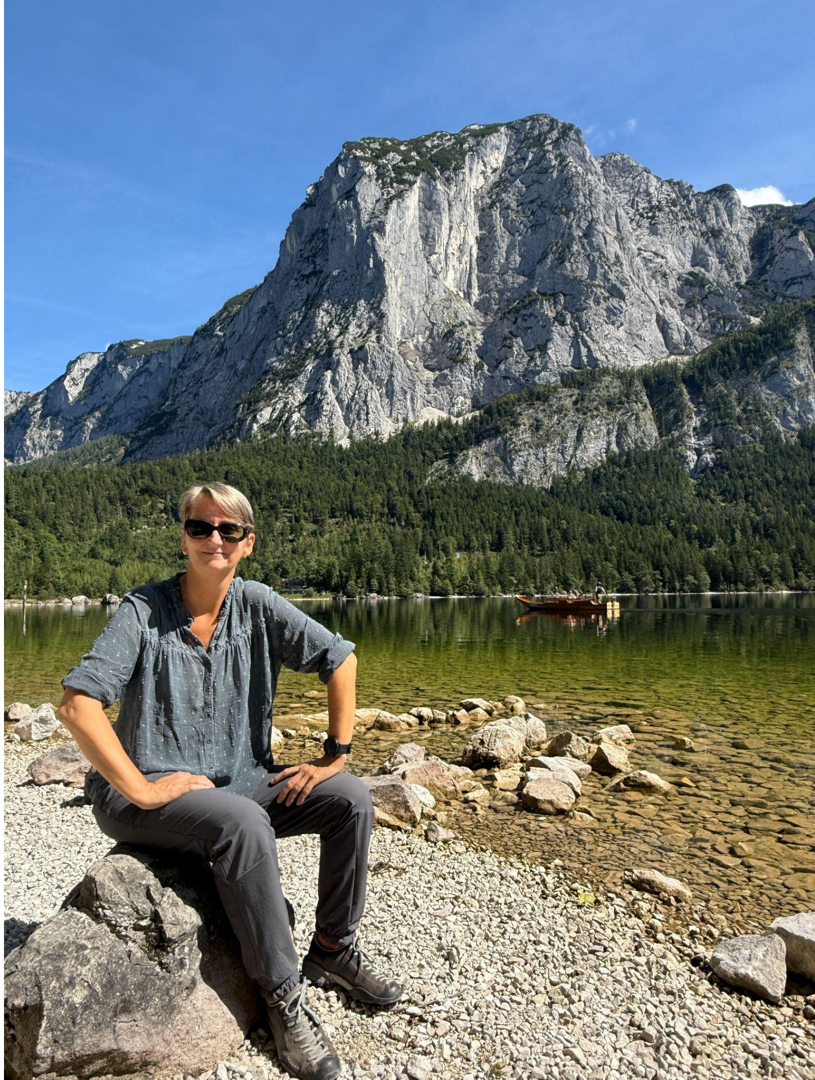
- Often perceived as less sensitive than textual data
- BUT: can reveal
  - identities
  - locations
  - social status
  - health conditions
  - religious affiliations
  - political positions

Examples: photographs of people, digitised manuscripts with personal annotations, images of rituals, visual documentation of field work

Images

Even when individuals are unnamed, visual identifiability can make anonymisation effectively impossible!

Work assignment (2 minutes)
Consider which elements in the image potentially enable identification.



Images and Informed consent forms

- If signed informed consent form available, image can be published or reused
  - In practice, not always sufficient – may also be an ethical component

- *Informed consent was topic of fifth webinar of this series – if you missed it and want to learn more, slides and recording are accessible here:* *https://ark.dasch.swiss/ark:/72163/1/0810/5f0JdGIgSpSn80d97h5aGgX*

Informed consent forms can

- Specify intended uses, such as publications, teaching, exhibitions, or online dissemination
- Distinguish access options
- Provide a legal and ethical framework for reuse and long-term preservation

Informed consent forms cannot

Eliminate all ethical risks:

- Participants may not fully anticipate future forms of reuse or technological change
- Images may gain new meanings or sensitivities over time
- Consent may be given under social, economic, or institutional pressure

Consent should be contextual and revisable!

Images: What should be avoided without informed consent

Faces and identifiable bodies

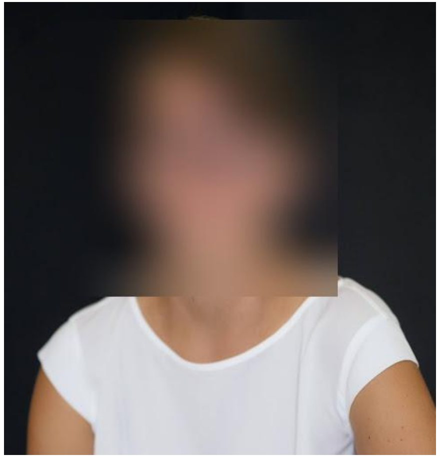
- Blurring faces is a common solution
  - but body posture, clothing, tattoos, or context may still allow identification!

Sacred or restricted images

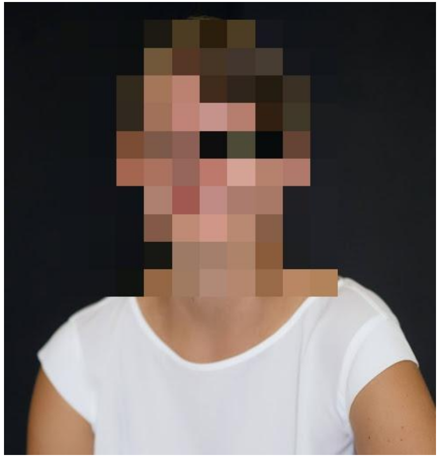
- Sacred objects, rituals, burial sites, community-restricted knowledge
  - Anonymisation not an ethical solution – restricted access, limiting reuse rights or refraining from publication altogether may be!

Images: Blurring faces

Was topic of first webinar of this series – if you missed it and want to learn more, python code, slides, and recording are accessible here: [*https://ark.dasch.swiss/ark:/72163/1/0810/4tzTlf7OSI2wAV2oKblwiQc*](https://ark.dasch.swiss/ark:/72163/1/0810/4tzTlf7OSI2wAV2oKblwiQc)



blurring

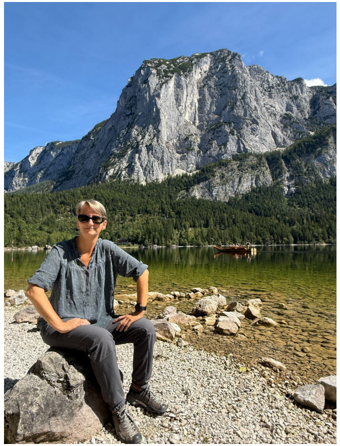


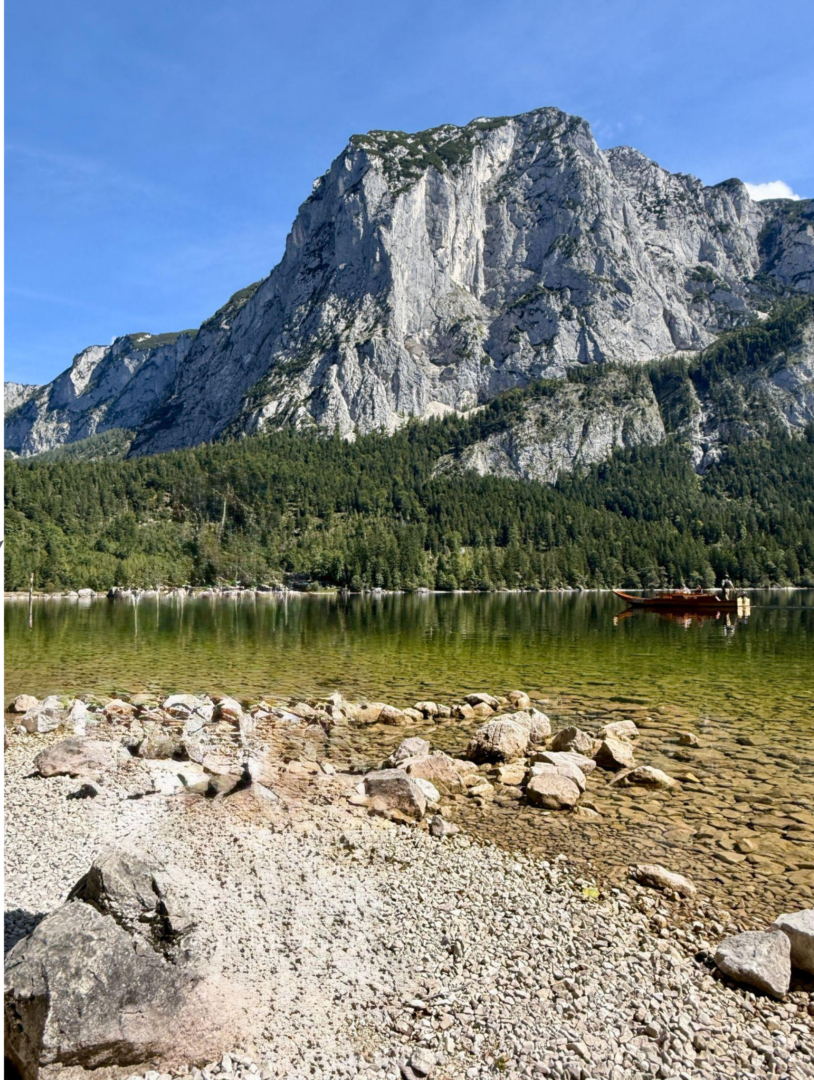
pixelation

Attention: image deconvolution may be / is possible!

Images: Remove people

- If person(s) on image are not the relevant thing – see if AI features within image processing software can help you

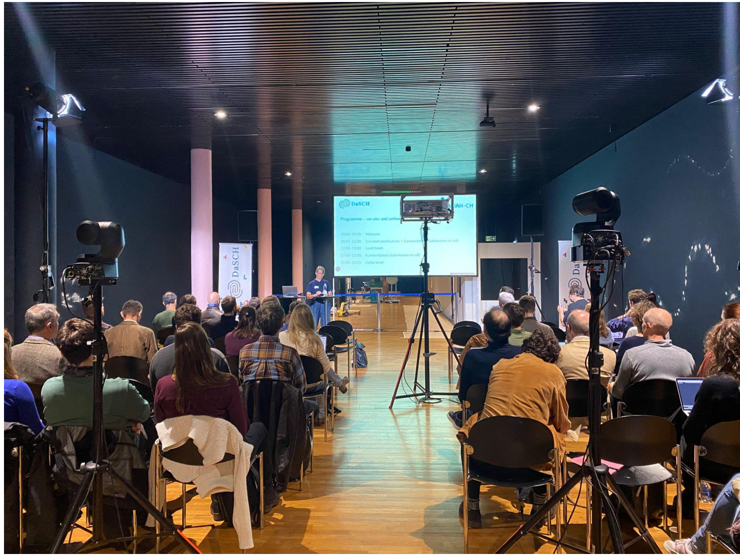




Images: What is (sometimes) allowed

Best practices: Identifiability

1. People seen from the back, at a distance, or as part of a crowd, or visually incidental to the main subject
   - ○ Usually considered low-risk from data protection perspective
   - ○ Usually do not allow direct identification
   - ○ *low identifiability does not automatically mean ethical neutrality*!



Images: What is (sometimes) allowed

Best practices: Identifiability

2. Incidental presence versus research focus
    - key distinction!
    - publication often ethically acceptable if merely incidental, e.g.if
        - image was taken in a public space
        - no sensitive activity is depicted
        - no attempt is made to interpret or categorise the individuals

Images: What is (sometimes) allowed

Best practices: Identifiability

3. People of public interest
    - politicians, artists, activists, or religious leaders
    - treated differently legally and ethically
    - But …
        - Public interest does not imply unlimited consent
        - Images should relate directly to the public role or activity
        - Private, intimate, or vulnerable moments remain ethically sensitive
        - scholarly purpose of using the image should be justified

Images: What is (sometimes) allowed

Archival images

- usually no explicit consent exists,
- individuals depicted are no longer alive,
- images may have been produced under historical power asymmetries

Common assumption:

- age alone removes all concerns
- only partly true
- retrospective ethical assessment is advisable

Images: What is (sometimes) allowed

Archival images: Best practices

- assess not only legal status (e.g. copyright expiry) but potential social or cultural sensitivity
- evaluating the original context of image production and providing contextual framing that acknowledges historical conditions
- avoid unnecessary identification when it adds no scholarly value
- consider whether descendants or communities may (still) be affected, whether the image reinforces harmful stereotypes or narratives

Images: Practical guidelines

Ask the following questions before publishing images

1. How easily could individuals be identified — directly or indirectly?
2. Is the person incidental or central to the image?
3. Does the image show people in a vulnerable, private, or sensitive situation?
4. Does publication serve a clear scholarly purpose?
5. Would anonymisation meaningfully reduce risk (or create false reassurance)?
6. Are there cultural, historical, or community-specific visibility norms to consider?

Ethnology & Indigenous Data

Ethnology & Indigenous Data

Ethnological and qualitative data often includes:

- Personal narratives
- Cultural knowledge
- Political or spiritual beliefs
- Community-internal information

Even when individuals are anonymised, contextual re-identification can be easy — especially in small or marginalised communities!

Indigenous Data

Indigenous and local communities

- anonymisation may be the wrong frame altogether, because
  - Collective rights, not just individual consent
  - Data sovereignty
  - Cultural protocols governing access, reuse, and interpretation

*Frameworks such as the **CARE** (Collective Benefit, Authority to Control, Responsibility, Ethics) **principles** exist. Were topic of second webinar of this series – if you missed it and want to learn more, slides, and recording are accessible here: https://ark.dasch.swiss/ark:/72163/1/0810/j5q2Rw9tTbKVaufeXGR=7Qi*

Indigenous and Community-Based Data: Best practices

- Prioritise community consent, not just individual consent
- Allow communities to define:
  - What may be shared
  - With whom
  - Under which conditions

FORS Data and services for the social sciences

DaSCH

Useful tools for anonymisation and pseudonymisation

Useful tools

- A wide range of tools can support anonymisation of qualitative and quantitative data in SSH
- For practical guidance, consult FORS selection of tools
    - https://forscenter.ch/wp-content/uploads/2026/02/a-practical-guide-to-anonymisation-tools.pdf

FORS Data and services for the social sciences

DaSCH

Save the Date!

	Seventh Webinar
Topic	Reproducibility, Linked Open Data & Semantic Interoperability
When	May 21, 2026 at 2pm
Where	Online

Source: Freepik, Marsiholo

FORS Data and services for the social sciences

DaSCH

Questions?

Source: Abscent84 / Getty Images