Ethical Open Data Principles

Share and care for your data, and don’t forget to anonymise

Marc Schulder
Universität Hamburg

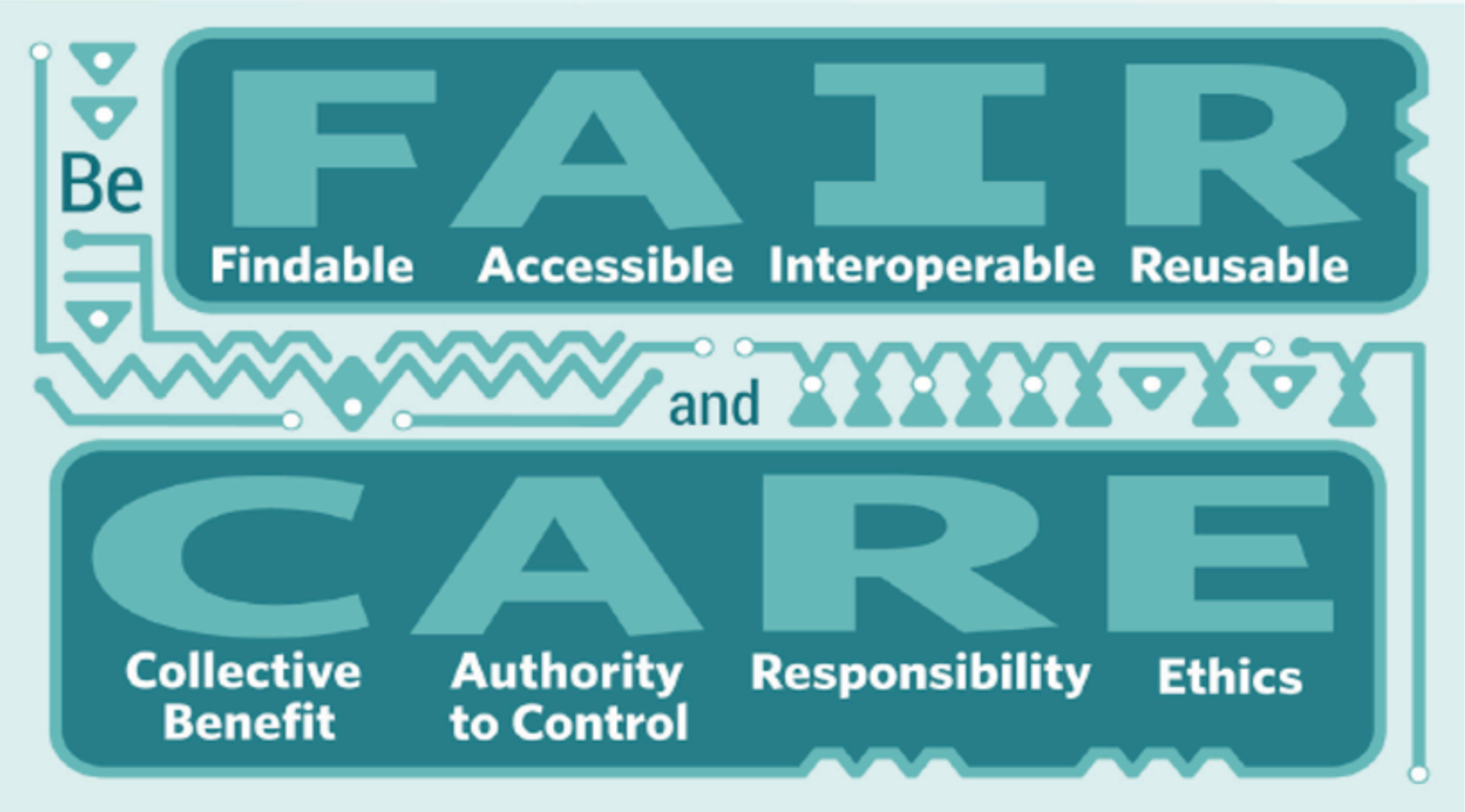
EASIER Autumn School 2023

Questions

- Can I publish my data? Should I publish my data?
- What metadata do I collect? What do I publish?
- What does informed consent entail?
- Do I need a licence?
- What and how do I anonymise?
- How do I make sure my corpus is long-lived?

https://www.go-fair.org

(Wilkinson et al., 2016)



https://www.gida-global.org/care

(RDA International Indigenous Data Sovereignty IG, 2019)

How to be FAIR when you CARE: […], Schulder & Hanke (2022)

FAIR Principles

- **Findable:** Data should be easy to find for both humans and machines.
- **Accessible:** Users need to know how to access (meta)data.
- **Interoperable:** Follow open standards to be compatible with tools and other data.
- **Reusable:** Data should be clearly documented and licensed.

CARE Principles

applied to deaf communities

- **Collective Benefit:** How do communities benefit from the work?
- **Authority to Control:** Participants should retain control of their data.
- **Responsibility:** Document how data is used and how it benefits deaf community. Set usage conditions that protect participants and deaf community in general.
- **Ethics:** The deaf communities’ rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

Data Collection

- Purpose of data collection
- Informed Consent
- Data Sharing
- Content Approval
- Collection of metadata

Data Publication

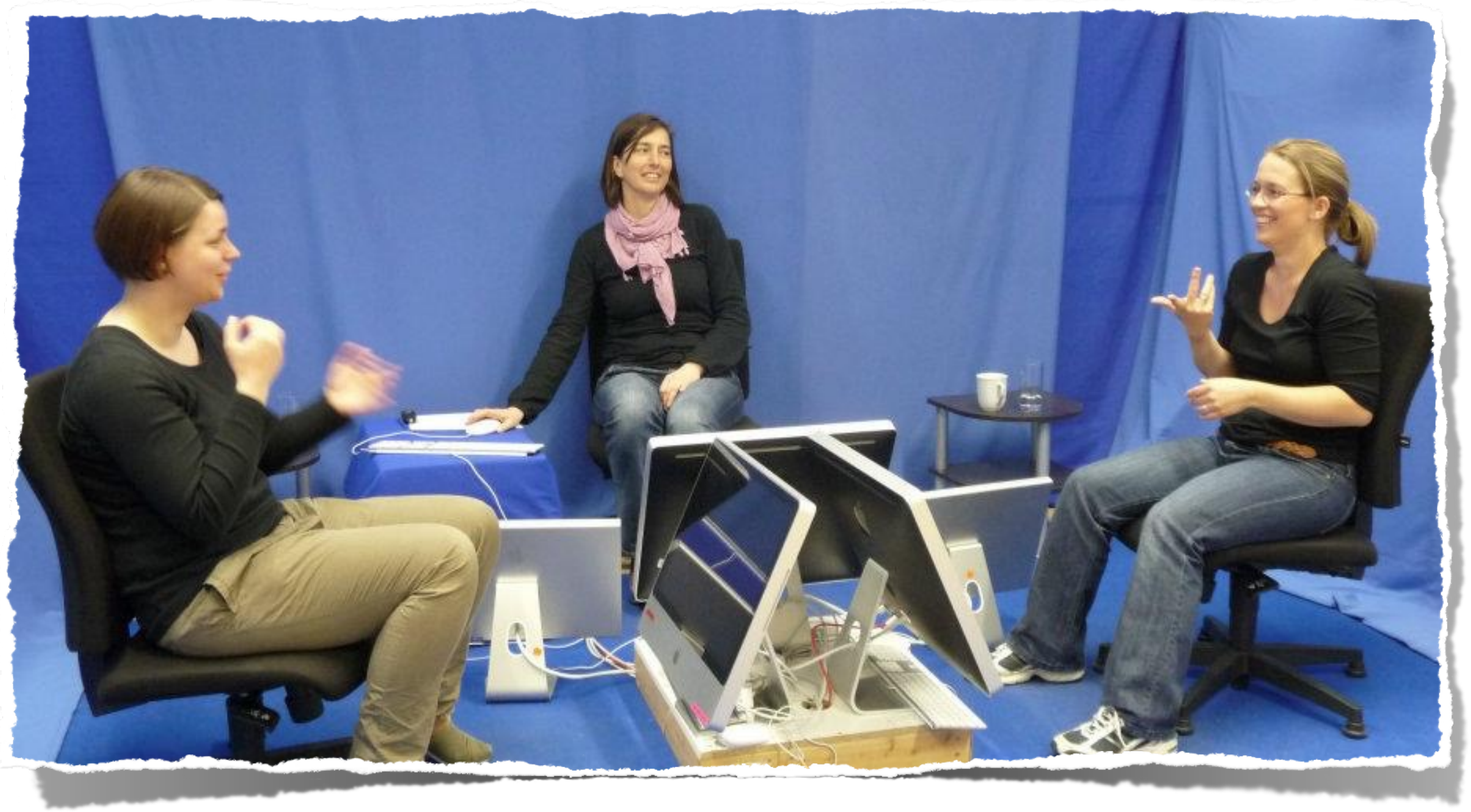
- Licensing data usage
- Anonymisation
- Persistent identifiers
- Metadata
- Data storage

Purpose of Data Collection

Be FAIR Findable Accessible Interoperable Reusable and CARE Collective Benefit Authority to Control Responsibility Ethics

- **DGS Corpus:**
  - Reference corpus of DGS for linguistic research
  - Corpus-based dictionary
  - Cultural heritage

DGS-KORPUS

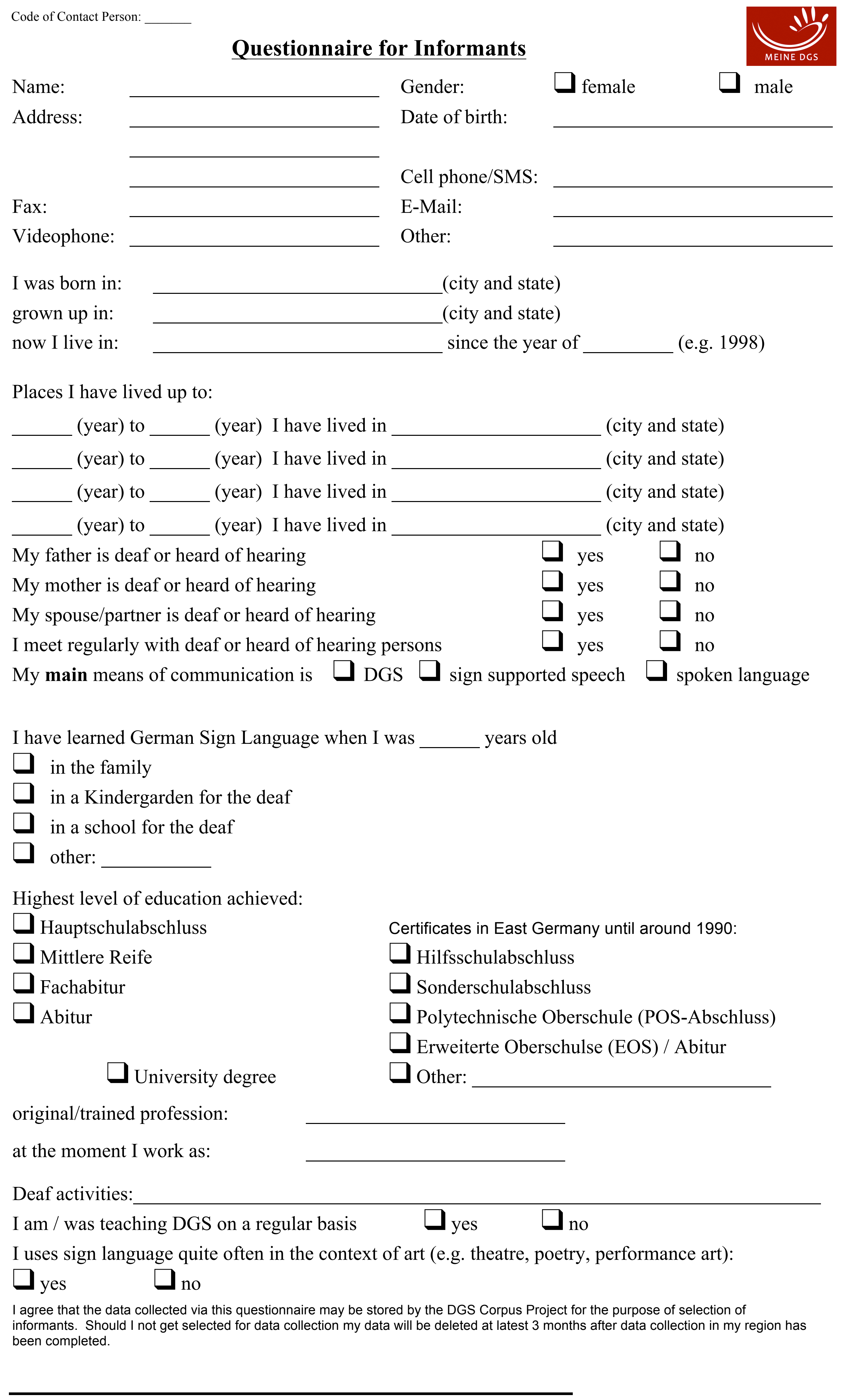


Informed Consent

- **Information (Written & Signed):**
  - Purpose of research
  - Information collected (video, contact info, metadata)
  - Use of data within project
  - Data publication & sharing
- **Consent:**
  - For what uses can my data be published/shared? (Digital heritage, teaching, linguistics, NLP?)
  - Let other researchers contact me
  - Allow me to approve content (incl. partial exclusions)
  - Allow me to change my mind in future

FANTASTIC METADATA
AND HOW TO COLLECT IT

- **Participant selection:**
  - Basic information for balancing dataset
  - Avoid collecting information you don’t need for selection
- **Dataset metadata:**
  - Detailed information that might become important in future research
- **Published metadata:**
  - Very general information that does not impact participant’s privacy



Code of Contact Person: _______

Questionnaire for Informants

Name: ____________________ Gender: ❑ female ❑ male
Address: ____________________ Date of birth: ____________________
____
____________________ Cell phone/SMS: ____________________
Fax: ____________________ E-Mail: ____________________
Videophone: ____________________ Other: ____________________

I was born in: ______________________(city and state)
grown up in: ______________________(city and state)
now I live in: ______________________ since the year of _________ (e.g. 1998)

Places I have lived up to:
______ (year) to ______ (year) I have lived in ________________ (city and state)
______ (year) to ______ (year) I have lived in ________________ (city and state)
______ (year) to ______ (year) I have lived in ________________ (city and state)
______ (year) to ______ (year) I have lived in ________________ (city and state)
My father is deaf or heard of hearing ❑ yes ❑ no
My mother is deaf or heard of hearing ❑ yes ❑ no
My spouse/partner is deaf or heard of hearing ❑ yes ❑ no
I meet regularly with deaf or heard of hearing persons ❑ yes ❑ no
My **main** means of communication is ❑ DGS ❑ sign supported speech ❑ spoken language

I have learned German Sign Language when I was ______ years old
❑ in the family
❑ in a Kindergarden for the deaf
❑ in a school for the deaf
❑ other: ___________

Highest level of education achieved:
❑ Hauptschulabschluss
❑ Mittlere Reife
❑ Fachabitur
❑ Abitur
❑ University degree

Certificates in East Germany until around 1990:
❑ Hilfsschulabschluss
❑ Sonderschulabschluss
❑ Polytechnische Oberschule (POS-Abschluss)
❑ Erweiterte Oberschulse (EOS) / Abitur
❑ Other: ____________________

original/trained profession: ____________________

at the moment I work as: ____________________

Deaf activities:______________________________________
I am / was teaching DGS on a regular basis ❑ yes ❑ no
I uses sign language quite often in the context of art (e.g. theatre, poetry, performance art):
❑ yes ❑ no

I agree that the data collected via this questionnaire may be stored by the DGS Corpus Project for the purpose of selection of informants. Should I not get selected for data collection my data will be deleted at latest 3 months after data collection in my region has been completed.

____

Handbuch für Kontaktpersonen Teil I: Projekt, Werbung, Informantensuche, Raumsuche; König et al. (2020)

Data Collection

- Purpose of data collection
- Informed Consent
- Data Sharing
- Content Approval
- Collection of metadata

Data Publication

- Licensing data usage
- Anonymisation
- Persistent identifiers
- Metadata
- Data storage

Licensing Data Usage

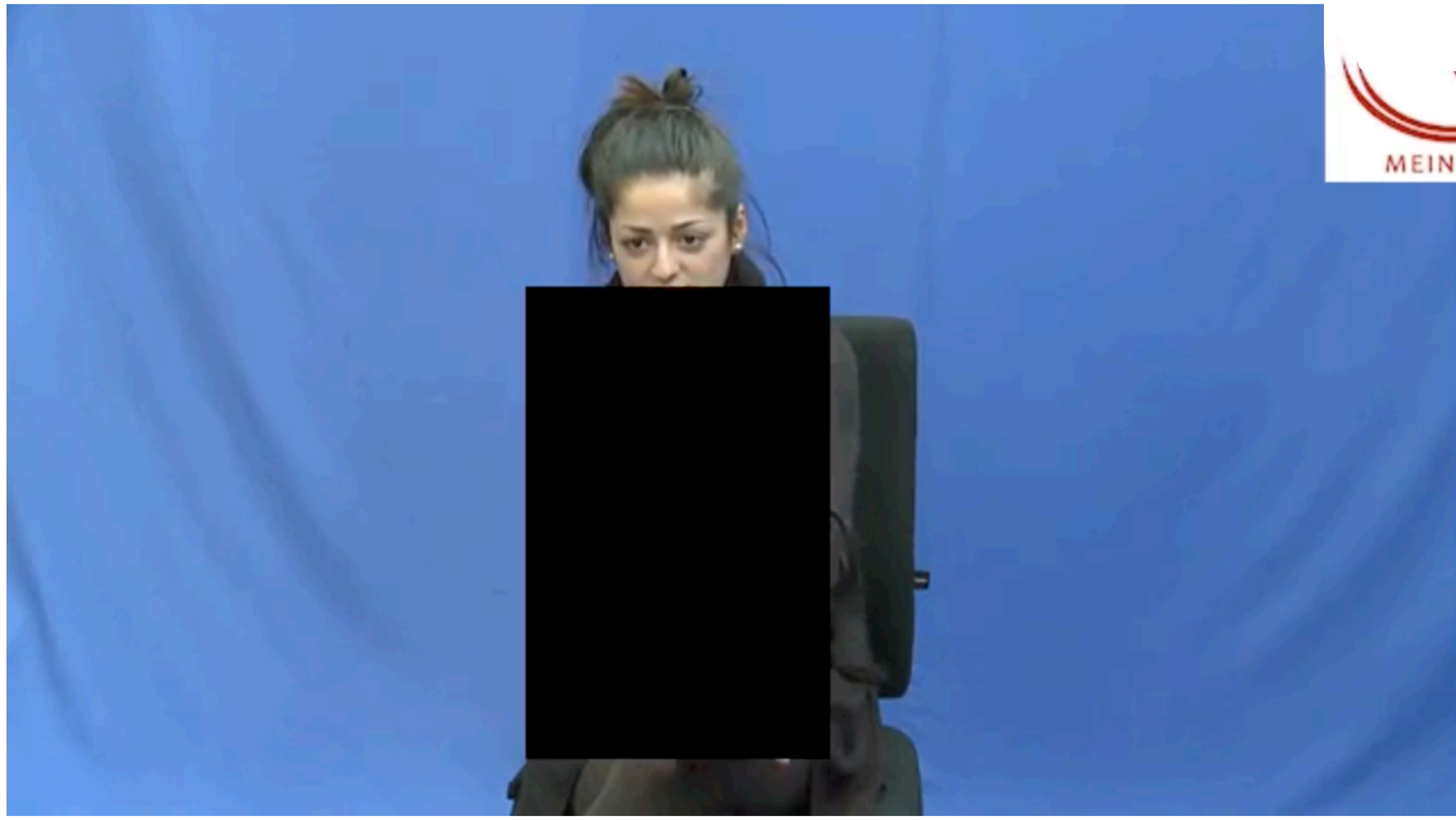
- **FAIR:**
  - Make your licence as open **as possible,** but as restrictive as **necessary**
- **CARE:**
  - Can your data cause harm by being publicly visible?
  - Does a permissible licence enable harmful uses of your data?
  - Are all uses permitted by the licence covered by your informed consent?

Licensing Data Usage

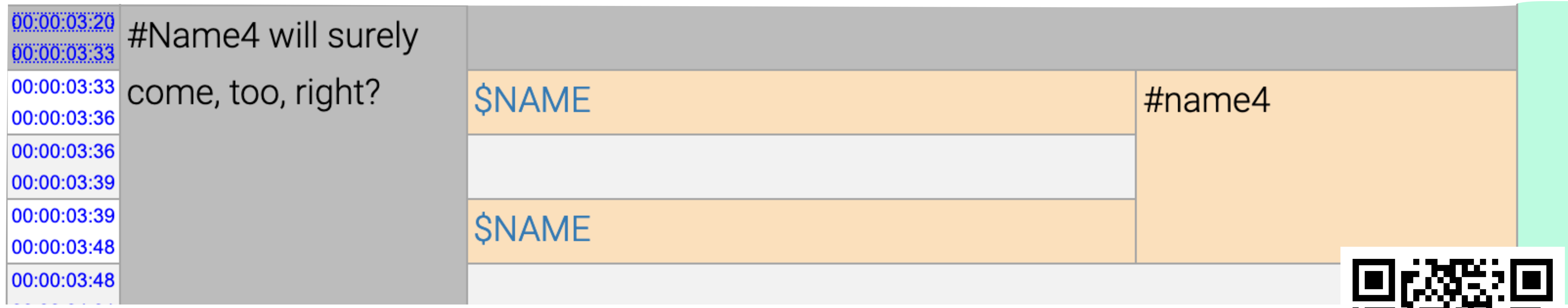
- **Who wants to use data?**
  - private/non-commercial/commercial entity
- **What for?**
  - Cultural heritage? Teaching? Linguistics? Machine learning?
  - view vs download
- **Which recordings?**
  - Personal story? Opinions? Retelling? Scripted?
- Does your **informed consent cover all uses?**
- **Same licence for all uses? Same licence for all data?**

Anonymisation

- **What content**
  - Excluded during content approval
  - Personal identifiable information
    - participants & 3rd parties
    - names, dates, exact locations


[redacted]

- **What data**
  - Metadata *(coarsen)*
  - Video *(blacken)*
  - Translations, glosses, mouthing *(categorisation)*


00:00:03:20
00:00:03:33
00:00:03:33
00:00:03:36
00:00:03:36
00:00:03:39
00:00:03:39
00:00:03:48
00:00:03:48

#Name4 will surely come, too, right?

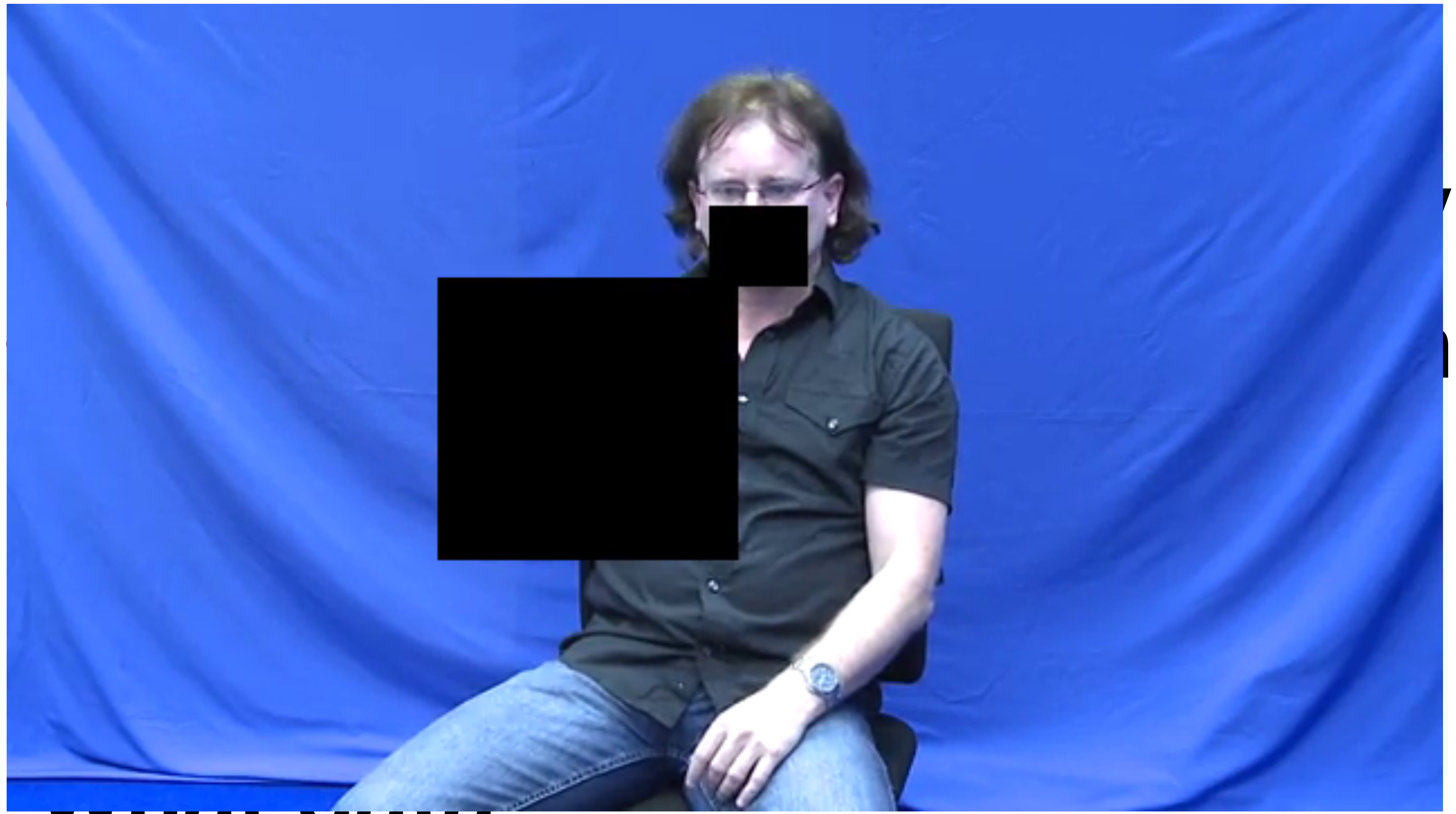
$NAME

#name4

$NAME

Approaches to the Anonymisation of Sign Language Corpora, Isard (2020)

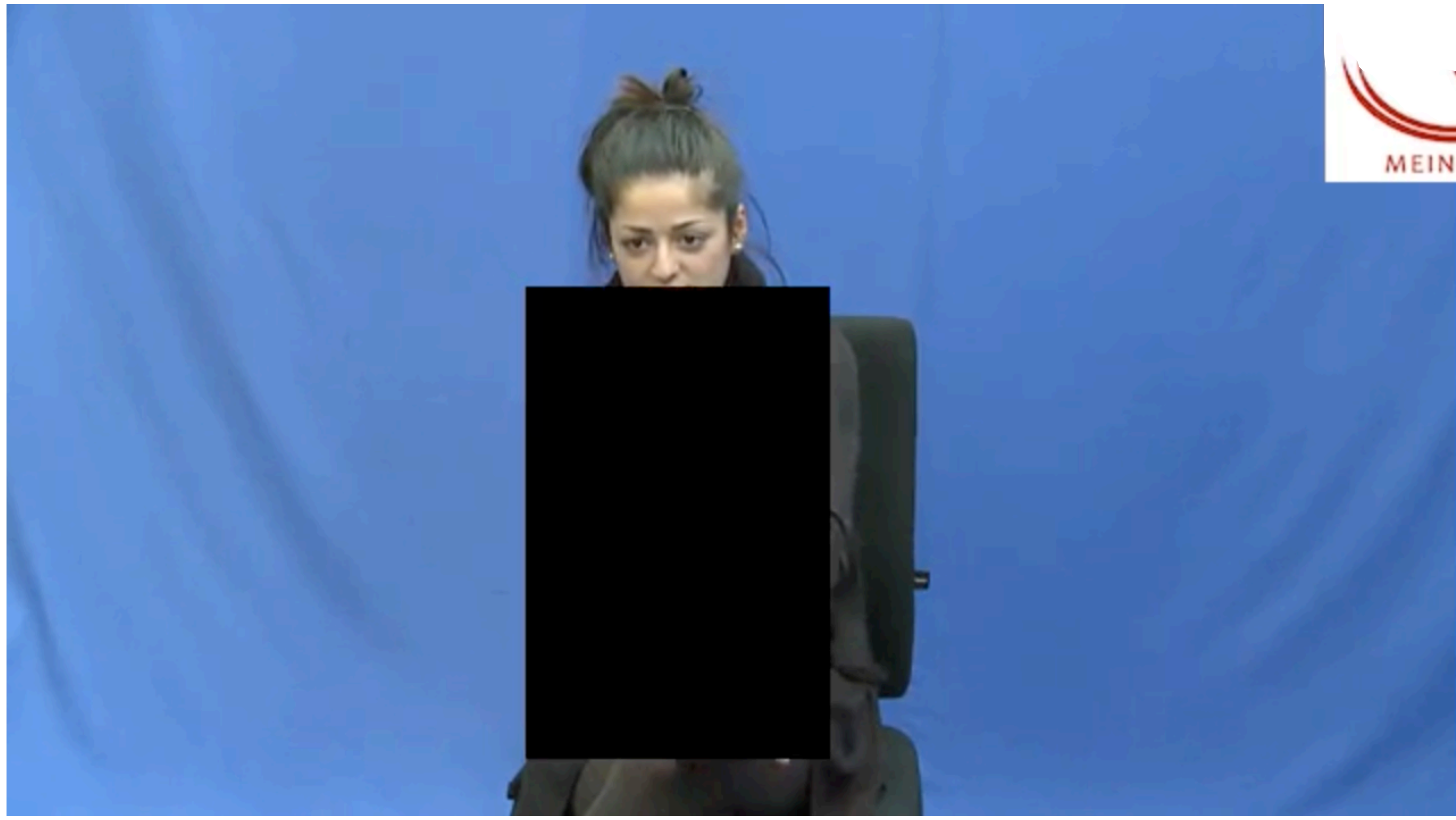
Anonymisation

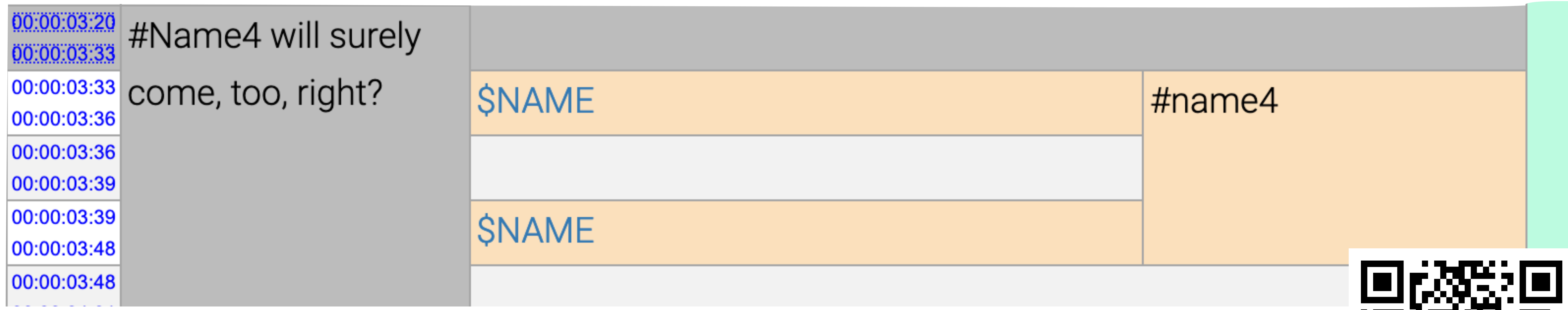


- al

  n

- What data
  - Metadata *(coarsen)*
  - Video *(blacken)*
  - Translations, glosses, mouthing *(categorisation)*





00:00:03:20 00:00:03:33	#Name4 will surely come, too, right?		
00:00:03:33 00:00:03:36		$NAME	#name4
00:00:03:36 00:00:03:39			
00:00:03:39 00:00:03:48		$NAME	
00:00:03:48			

Approaches to the Anonymisation of Sign Language Corpora, Isard (2020)

Persistent Identifiers

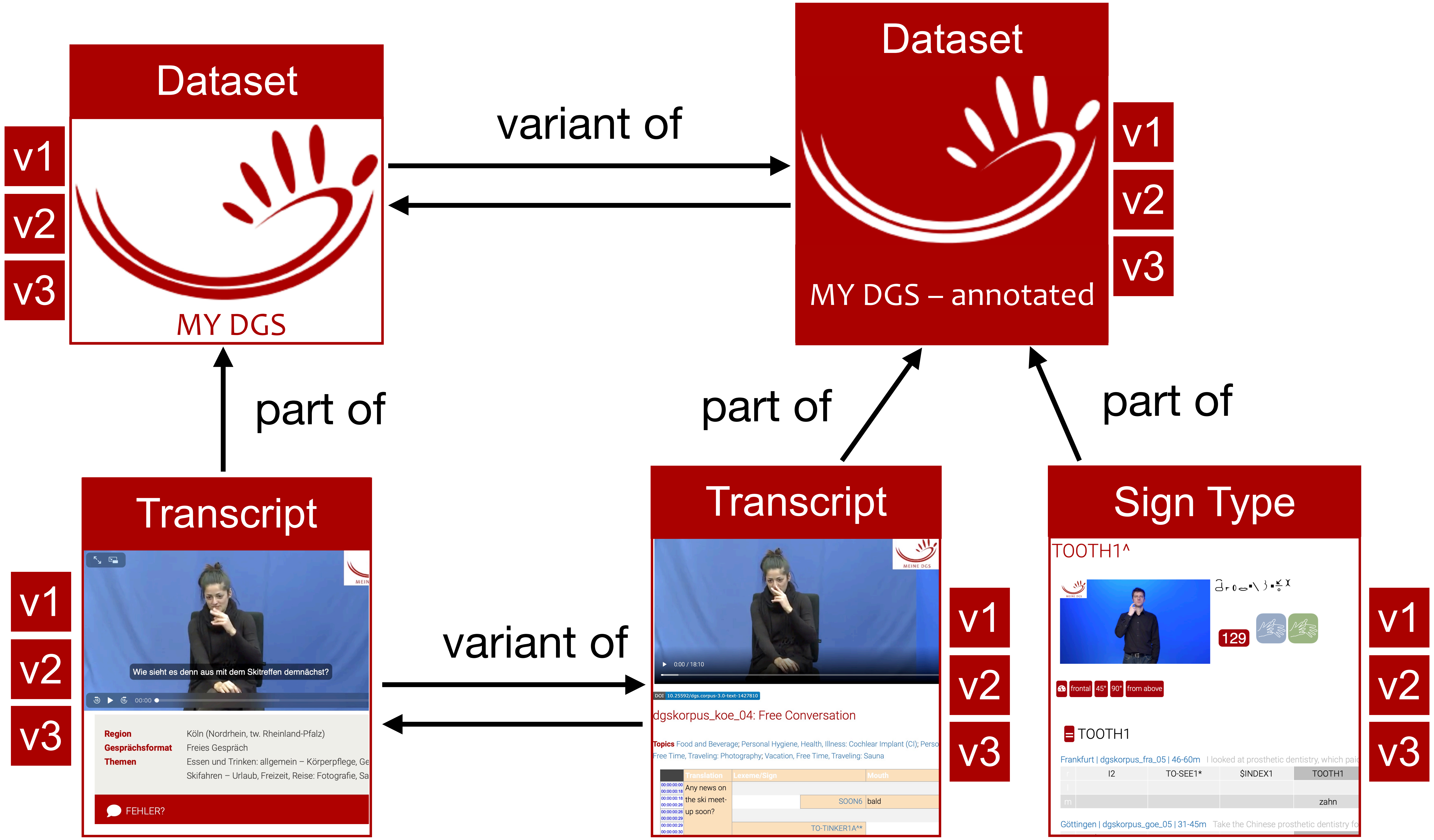
DOI 10.25592/dgs.corpus-3.0

- Dataset identifiers should be **unique** and **persistent**
  - **Persistent:** never change, unlike normal URL
  - **Unique:** Refer to one version of one entity
    - Different dataset variants (e.g. research vs heritage)? *Different ID!*
    - Changed something? *New version with new ID!*
- Good identifiers also have metadata:
  - Authors, title, description
  - References to related identifiers
    (version, variant, publication, child entity)

Persistent Identifiers

DOI 10.25592/dgs.corpus-3.0

Be FAIR and CARE
Findable Accessible Interoperable Reusable
Collective Benefit Authority to Control Responsibility Ethics


Dataset
MY DGS
v1
v2
v3
variant of
Dataset
MY DGS – annotated
v1
v2
v3
part of
part of
part of
Transcript
Wie sieht es denn aus mit dem Skitreffen demnächst?
Region Köln (Nordrhein, tw. Rheinland-Pfalz)
Gesprächsformat Freies Gespräch
Themen Essen und Trinken: allgemein – Körperpflege, Ge
Skifahren – Urlaub, Freizeit, Reise: Fotografie, Sa
FEHLER?
v1
v2
v3
variant of
Transcript
dgskorpus_koe_04: Free Conversation
Topics Food and Beverage; Personal Hygiene, Health, Illness: Cochlear Implant (CI); Perso
Free Time, Traveling: Photography; Vacation, Free Time, Traveling: Sauna
Any news on the ski meet-up soon?
SOON6 bald
TO-TINKER1A*
v1
v2
v3
Sign Type
TOOTH1^
129
frontal 45° 90° from above
TOOTH1
Frankfurt | dgskorpus_fra_05 | 46-60m I looked at prosthetic dentistry, which pa
I2 TO-SEE1* $INDEX1 TOOTH1
zahn
Göttingen | dgskorpus_goe_05 | 31-45m Take the Chinese prosthetic dentistry
v1
v2
v3

Metadata

Be FAIR Findable Accessible Interoperable Reusable and CARE Collective Benefit Authority to Control Responsibility Ethics

- **DOI Metadata**
  - Generic dataset information (title, authors, description,…)
  - Qualified references to other versions/datasets/publications/…
- **CMDI Metadata**
  - Linguistic information (participant metadata, genre, context, content languages, annotation languages)

Persistent Identifiers and Metadata for the Public DGS Corpus, Hanke (2021)

- **Documentation**
  - Project reports
  - For (computational) re-use: Data Statement

Bender & Friedman (2018)

Data Storage

- **Backup** strategy for in-progress work
- **Archive** original data & published works
- **Data Custodian**: Who is in charge of data in X years time?
- Storage **location** important for sensitive data
    - What country? What organisation controls data?
    - Archive should be FAIR:
        - **persistent identifiers** (e.g. DOI)
        - public dataset **metadata**
        - allow public and restricted **access**

How to be FAIR when you CARE: […], Schulder & Hanke (2022)

Discussion

Thanks to Sabrina Wähl who co-wrote a previous version of this talk with me