

CLARIN2020 Bazaar Session: Sharing corpora of disordered speech and finding relevant use cases

Moderators: Henk van den Heuvel & Esther Hoorn

Wednesday 7 October, 14.50 - 15.50

DELAD initiative



- Initiative to collect and share corpora of speech with disorders (CSD):
<http://delad.net/>
- Partners are a mix of researchers, infrastructure specialists, legal experts
- DELAD organises annual workshops since 2015 where these groups convene
- Since 2017 under CLARIN header and support

Topics addressed:

- Examples of CDS
- Guidelines for collecting and sharing CDS
- Ethics and legal aspects
- Levels of anonymisation
- Layered access of data
- Integration of CDS in the CLARIN infrastructure
- Formats
- Relevant metadata

DELAD initiative



- Collaboration with CLARIN Knowledge Centre for Atypical Communication Expertise
- For data storage hosting and sharing DELAD cooperates with ACE:
 - The Language Archive at MPI Nijmegen: <https://archive.mpi.nl/tla/>
 - Talkbank at CMU: <https://talkbank.org/>
- Use case: <https://phonbank.talkbank.org/access/Clinical/PCSC.html>
- Next DELAD workshop: 25-29 January 2021, online
 - Access options for CSD
 - Space for researchers to present their datasets and solutions/wishes for sharing them
 - GDPR-issues & DPIA, role play

LREC 2020 paper: Corpora of Disordered Speech in the Light of the GDPR: Two use cases from the DELAD Initiative



Needed for research purposes:

- appropriate safeguards which ensure technical and organisational measures
- legal ground and compliance with data protection principles
- privacy by design
- open question: who is controller in research collaboration...shared responsibilities; GDPR says: clarify & be transparent
- derogations for research not harmonized, but still need to be in keeping with the essence of data protection and discipline-specific good practices

How to find appropriate measures in innovative use cases, that want to share sensitive data?

Fraunhofer publica:

The Data Protection Impact Assessment
According to Article 35 GDPR

A Practitioner's Manual

<http://publica.fraunhofer.de/documents/N-590015.html>

Why DPIA in research:

- high risks (derogations, open science)
- shared responsibilities
- German method
(protection goals, standard data protection model, multistakeholder approach)



COMENIUS project: Privacy in Research: asking the right questions



Website:

<https://sites.google.com/rug.nl/privacy-in-research/home>

- active learning
 - online course
 - roleplay game
- real cases
- context of teaching
 - multidisciplinary methods,
 - Ethics, Research Integrity
 - and data management

What could the DPIA method bring for CDS data?

-applied on use cases of the ELRA article

Use Case 1:

Keywords: clinical datasets, different sources, multidisciplinary effort in collection, sensitive health data, secure system, ethical assessment, transparency on data sharing with vulnerable data subjects

-design appropriate measures
-the concept of a reference DPIA

1. design and awareness for importance of collaboration (legal, IT, researcher, data subjects, third parties)

Use Case 2:

Keywords: archival recordings, metadata of hearing impaired children, legal ground, further processing, sensitive nature of metadata, balancing assessment, stored at TLA

2. documented measures in report is also way to consult & get green light of the authority

-joint responsibilities: promote documentation: develop (ethical) standards.

A holistic multistakeholder approach: Example Parkinson Case

<https://youtu.be/H3WFEfieLm4>



extra:

- responsibilities of researcher as developer
- algorithmic transparency
- ethics by design
- understand diverse forms of regulation: other legal perspectives (IP, health apps)

January event:

a chance to get experts from the research community involved

Conclusion and way forward



1. training next generation with holistic approach
2. explore integration of institutional responsibility and discipline specific expertise
3. develop transparency & standards on shared responsibilities in research context

Interested?

Role Card

Researcher



Your role

Researchers to be involved are, for instance:

- ✓ A primary investigator (PI), because formally the PI will often be responsible
- ✓ An applicant in a project in which awareness of data protection issues needs to be documented
- ✓ A PhD student who takes the daily decisions about the management of personal data
- ✓ A Master's student who is developing their research plan
- ✓ ...

Why we need you in the team:

- ✓ Building data protection in the research design requires an active involvement of the researchers
- ✓ To help to design the dataflow and evolve your data management plan
- ✓ To identify good practices to deal with personal data in your research field
- ✓ To analyse how the proposed technical and organizational solutions fit into your research process
- ✓ ...

Get Started

This card will help you to ask the right questions

Protection Goals

IT security goals:

1. Confidentiality
2. Integrity
3. Availability

Data privacy goals:

4. Unlinkability
5. Intervenableity
6. Transparency



Examples of questions*

1. Can people expect this to happen, even if they do not read the information I provide them with (i.e. the processing of secondary data)? Do I target the impact on people's fundamental rights, freedoms, and interests and not only on the risks to my organization?
2. Is it easy for people to exercise their rights to access, rectification, erasure etc.?
3. Did I identify all of the purposes of processing data in the data life cycle? Are all purposes compatible with the initial purpose?
4. In case I want to make the data 'open', how do I manage the risk of re-identification?
5. Are there data items I could remove (or hide) without compromising the purpose of the process?
6. How long do I need to store the data? For which purpose(s)? Are there discipline specific criteria for storage periods?

Question for Bazaar:

Are you interested in this approach for your use cases?

Like to participate in the role game in January?

