Recent developments at ACE / DELAD for support of sharing CDS

DELAD Steering Group:
Nicola Bessell, *University College Cork*
Katarzyna Klessa, *Adam Mickiewicz university in Poznan*
Alice Lee, *University College Cork*
Satu Saalasti, *University of Helsinki*
Henk van den Heuvel, *Radboud University, Nijmegen*

1. Introducing DELAD
2. CDS annotation tools and techniques
3. Guidelines for consent & storage
4. DELAD DPIA Roleplay material
5. A walk through the DELAD website

DELAD CLARIN

DELAD Introduction

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

DELAD Introduction

- Goals of DELAD
- Provide a GDPR compliant platform for digital archving and sharing of disordered speech data, allowing for different levels of access
  - Audio
  - Visual
  - Transcription
  - Annotation
  - Instrumental data (acoustic, ultrasound, Mri, EPG etc.)
  - Meta data
  - Control data
- Enhance research and teaching through this resource
- Support development and assessment of therapeutic practice
- Networking and knowledge development through workshops
- Website with information, links

DELAD Introduction

- CLARIN K-Centre for Atypical Communication Expertise (ACE; https://ace.ruhosting.nl
- The Language Archive (TLA) at the Max Planck Institute for Psycholinguistics (https://tla.mpi.nl)
- TalkBank (https://www.talkbank.org)
- TAPAS (Training Network on Automatic Processing of PAthological Speech https://www.tapas-etn-eu.org/)
- ELRA (European Language Resource Association www.elra.info)

Recent developments at ACE / DELAD for support of sharing CDS

1. Introducing DELAD
2. CDS annotation tools and techniques
3. Guidelines for consent & storage
4. DELAD DPIA Roleplay material
5. A walk through the DELAD website

DELAD CLARIN

CDS annotation tools and techniques

We are interested to discuss the experiences of the workshop participants regarding annotation tools and techniques

We inspected selected features of annotation tools, data formats and procedures that can be useful at various stages of

- designing,
- developing,
- processing,
- analysing

speech resources with a special focus on resources for disordered speech.

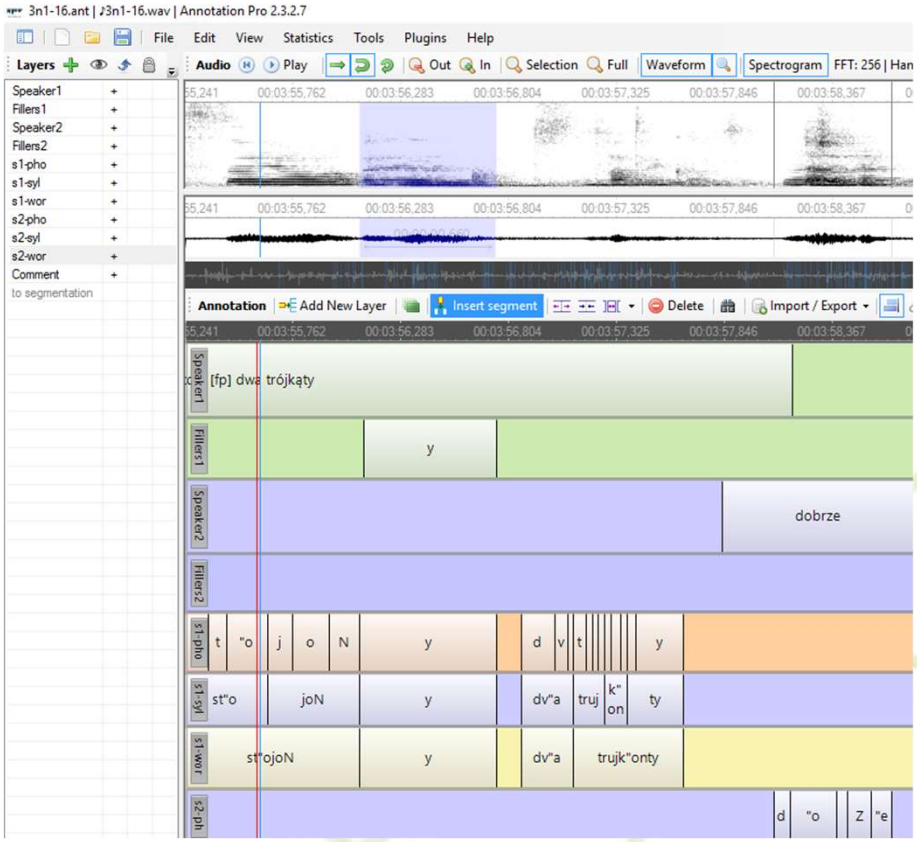
CDS annotation tools and techniques

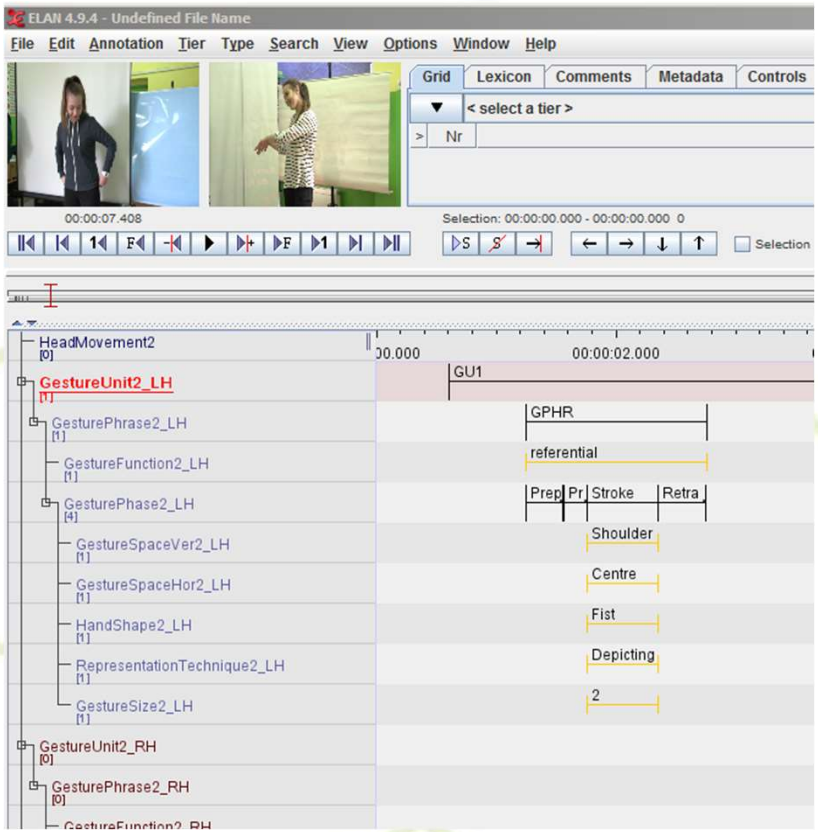
Requirements & examples

- **multilayer** annotations based on synchronized inputs representing different components of spoken communication (Praat, ELAN, Annotation Pro…)
  - the labelling schemes may involve both linguistic (e.g., phrases, words, syllables, individual sounds) and para/extralinguistic features (e.g., hesitation markers, physiological sounds produced by speakers or voice emotion correlates as well as gesture and mimicry annotation labels).
- **visual** representation of the multilayer annotations that are time-aligned with the speech signal display
- Extensions to the International Phonetic Alphabet - enable adding additional labels for **atypical** pronunciation (Ball, 2021; Ball et al., 2018)
- **uncertainty** labelling as an option in annotation tools, e.g. SPPAS (e.g. Bigi, 2021)

CDS annotation tools and techniques

Requirements & examples





Left: speech annotation in Annotation Pro, right: gesture annotation in ELAN. Illustrations from Borderland multimodal corpus (e.g. Karpiński & Klessa, 2018, Karpiński et al. 2018)),

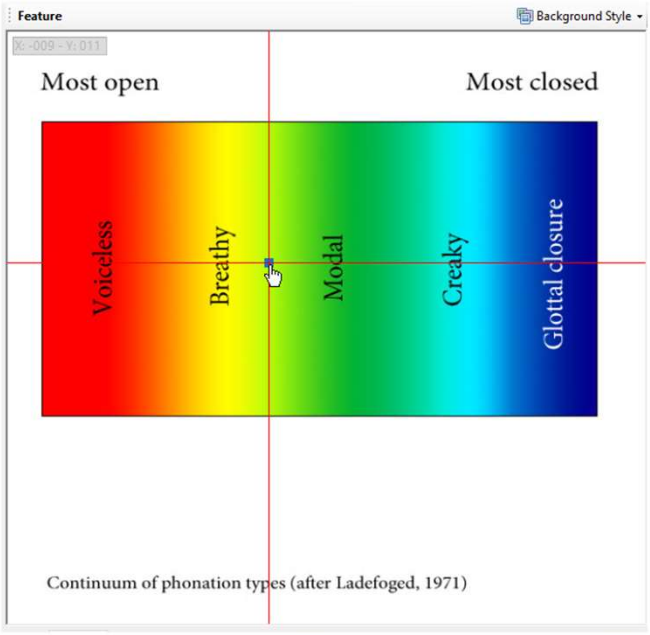
CDS annotation tools and techniques

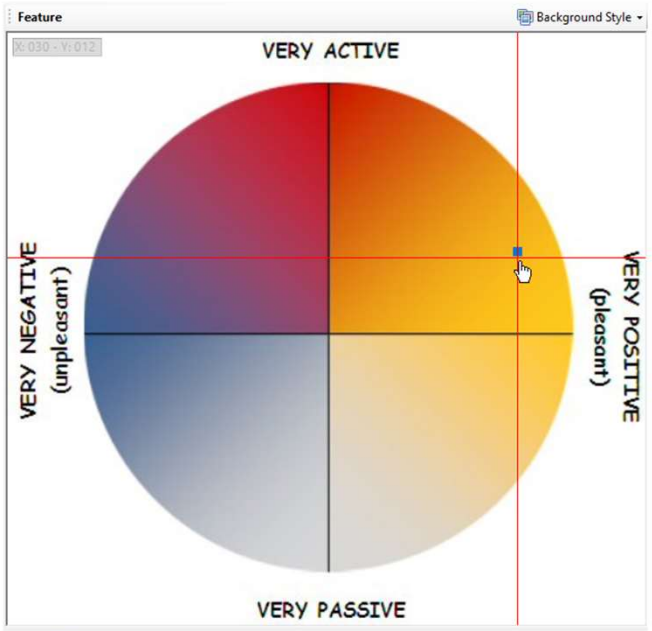
Discrete, continuous, and mixed rating scales

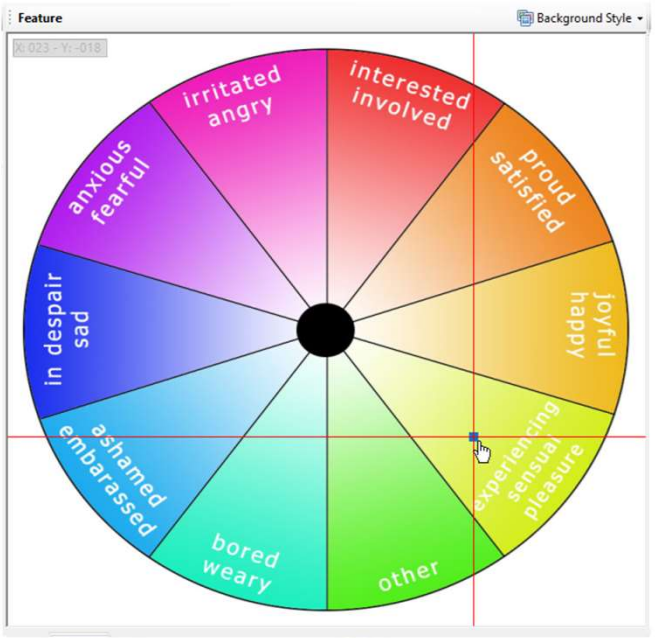
- When recordings are annotated using software tools, individual, **discrete** labels representing certain features are usually attached to subsequent segments (or boundary-delimited intervals).
  - Examples of such category-based labelling include labels denoting syllables, sounds, words, or whole phrases time-aligned with the speech sound signal.
  - Discrete labels are also used to annotate gestural behaviour and facial expressions, for example, to mark gesture types, gesture phases or functions (e.g., Ferré 2012; Jarmołowicz-Nowikow & Karpiński, 2011).
- However, when we consider the features of spontaneous or disordered speech, it sometimes turns out that discrete categories are difficult to apply.
- The profits of using **dimensional** rating scales are discussed with relation to e.g. voice quality, individual voice characteristics or speaker's states and attitudes.

CDS annotation tools and techniques

Discrete, continuous, and mixed rating scales







Examples of continuous and mixed rating scales available in the Annotation Pro software tool.

Interoperability of annotation data formats

- A common procedure of dealing with speech corpora, including CDS, is to **combine manual and automatic** ways of data processing. It also means using **different software tools** to process the same data sets.
  - For example, one can develop a corpus using one tool and analyse data with another one (cf. also Ide & Pustejovsky, 2010; Ide & Romary, 2007). See an example CDS analysis for French in B igii et al. (2015).
- It is therefore important to use **file formats that are readable** by different tools.
  - The annotation formats in ELAN, Annotation Pro and, SPPAS are XML-based, while Praat has its own file format but they all include time-stamp information. They can be converted to one another by means of either built-in functions or external converters (e.g., Annotation Pro enables quick data import and export between the tools).

Recent developments at ACE / DELAD for support of sharing CDS

1. Introducing DELAD
2. CDS annotation tools and techniques
3. Guidelines for consent & storage
4. DELAD DPIA Roleplay material
5. A walk through the DELAD website

DELAD CLARIN

Guidelines for consent & storage

- Info about research data management is available on websites of many universities & funded initiatives.
  - The amount of details vary.
  - The info is mostly generic – applicable to different disciplines.

- A webpage on consent & data storage on DELAD website: https://delad.ruhosting.nl/wordpress/guidelines-consent-storage/

Guidelines for consent & storage

- A few pointers to other useful websites e.g.,
  - UK Data Service (https://ukdataservice.ac.uk/)
  - National Coordination Point Research Data Management (https://www.lcrdm.nl/), The Netherlands

- A case scenario,
  - supported by a discussion of key ethics issues, &
  - a set of sample information sheet & consent forms.

Guidelines for consent & storage

The case scenario:

- *A researcher is planning to carry out a research project that investigates articulatory errors in adolescents and young adults with cerebral palsy using both auditory-perceptual and acoustic analyses. The data will also be compared to that of a group of age- and gender-matched typical speakers to examine whether there is any difference in speech characteristics between the two groups of speakers. Before speech data collection, the hearing and visual abilities of the participants will be screened to ensure that all have adequate abilities for taking part in the subsequent speech production tasks. …*

Guidelines for consent & storage

The case scenario (cont.):

- *… In addition, the participants with cerebral palsy will undertake a language test to document their language ability.The acoustic analysis of speech will be carried out by the research assistants of the project, whereas a group of typical individuals will be recruited as the listeners of the auditory-perceptual analysis of speech.The researcher also plans to have the speech data archived after the completion of this project to allow possible further research using the same set of data (e.g., analysis of voice quality) by other researchers and for education purposes in the future.*

Guidelines for consent & storage

- The case scenario captures common elements in many speech disorder research projects.
- Key ethical issues relevant to the case scenario.
- Researchers are reminded to consult their local policies & regulations to determine the relevance of our suggestions.
- We welcome further discussions to elaborate the example.

Guidelines for consent & storage

- Key ethical issues relevant to the case scenario:
  - Involvement of human participants
  - Collecting data from individuals under age 18
  - Collecting data from vulnerable individuals
  - Handling of personal information of participants
  - Handling of research data
  - Archiving & sharing research data with anonymised participant info
  - Data access level
  - Secondary analysis

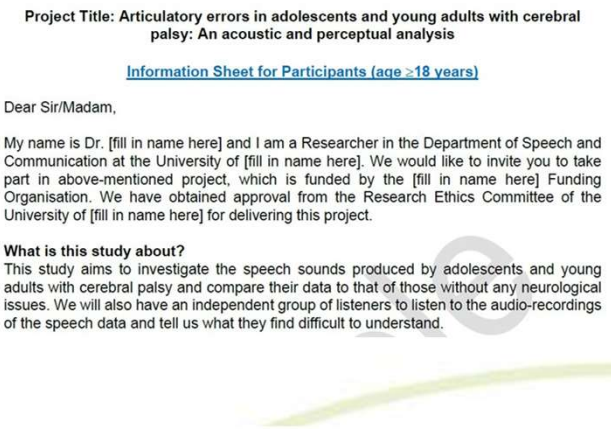
Guidelines for consent & storage

- Include a statement re archiving research data e.g.:
  - “DELAD stands for Database Enterprise for Language and speech Disorders (website: http://delad.net/) that aims to provide a channel for researchers to share corpora of speech of individuals with communication disorders with educators and researchers. DELAD has linked up with the Knowledge Centre for Atypical Communication Expertise (website: https://ace.ruhosting.nl/), a K-centre of CLARIN (Common Language Resources and Technology Infrastructure; website: https://www.clarin.eu/) for archiving and sharing the speech corpora through The Language Archive (website: https://archive.mpi.nl/tla/) and/or TalkBank (website: https://talkbank.org).”

Guidelines for consent & storage

Sample info sheets & consent / assent forms:

- For participants age ≥ 18
- For parents / guardians of participants age < 18
- For children participants age < 18
- For listeners



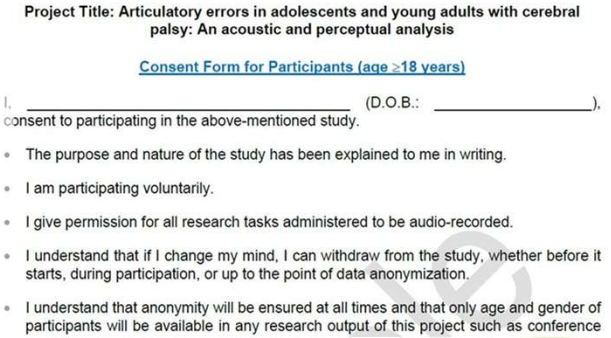
Project Title: Articulatory errors in adolescents and young adults with cerebral palsy: An acoustic and perceptual analysis

Information Sheet for Participants (age ≥18 years)

Dear Sir/Madam,

My name is Dr. [fill in name here] and I am a Researcher in the Department of Speech and Communication at the University of [fill in name here]. We would like to invite you to take part in above-mentioned project, which is funded by the [fill in name here] Funding Organisation. We have obtained approval from the Research Ethics Committee of the University of [fill in name here] for delivering this project.

What is this study about?
This study aims to investigate the speech sounds produced by adolescents and young adults with cerebral palsy and compare their data to that of those without any neurological issues. We will also have an independent group of listeners to listen to the audio-recordings of the speech data and tell us what they find difficult to understand.

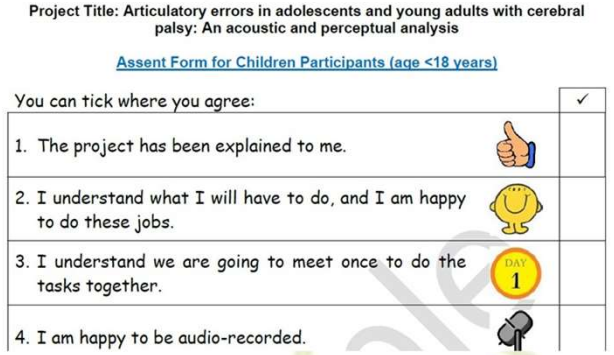


Project Title: Articulatory errors in adolescents and young adults with cerebral palsy: An acoustic and perceptual analysis

Consent Form for Participants (age ≥18 years)

I, _____ (D.O.B.: _____), consent to participating in the above-mentioned study.

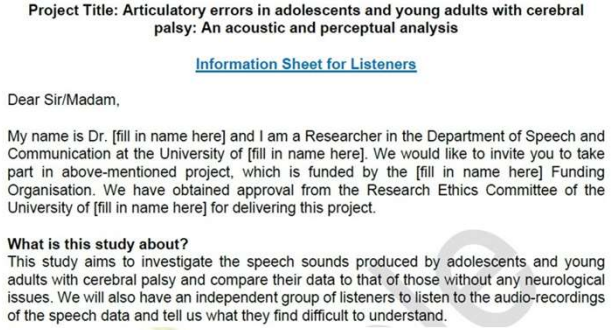
- The purpose and nature of the study has been explained to me in writing.
- I am participating voluntarily.
- I give permission for all research tasks administered to be audio-recorded.
- I understand that if I change my mind, I can withdraw from the study, whether before it starts, during participation, or up to the point of data anonymization.
- I understand that anonymity will be ensured at all times and that only age and gender of participants will be available in any research output of this project such as conference



Project Title: Articulatory errors in adolescents and young adults with cerebral palsy: An acoustic and perceptual analysis

Assent Form for Children Participants (age <18 years)

You can tick where you agree:	✓
1. The project has been explained to me.	
2. I understand what I will have to do, and I am happy to do these jobs.	
3. I understand we are going to meet once to do the tasks together.	
4. I am happy to be audio-recorded.	



Project Title: Articulatory errors in adolescents and young adults with cerebral palsy: An acoustic and perceptual analysis

Information Sheet for Listeners

Dear Sir/Madam,

My name is Dr. [fill in name here] and I am a Researcher in the Department of Speech and Communication at the University of [fill in name here]. We would like to invite you to take part in above-mentioned project, which is funded by the [fill in name here] Funding Organisation. We have obtained approval from the Research Ethics Committee of the University of [fill in name here] for delivering this project.

What is this study about?
This study aims to investigate the speech sounds produced by adolescents and young adults with cerebral palsy and compare their data to that of those without any neurological issues. We will also have an independent group of listeners to listen to the audio-recordings of the speech data and tell us what they find difficult to understand.

Recent developments at ACE / DELAD for support of sharing CDS

1. Introducing DELAD
2. CDS annotation tools and techniques
3. Guidelines for consent & storage
4. DELAD DPIA Roleplay material
5. A walk through the DELAD website

DELAD CLARIN

DELAD DPIA Roleplay material

During our previous workshop Esther Hoorn guided us through a DPIA (Data Protection Impact) Roleplay which she adapted for the DELAD community. Esther is a member of CLIC.

The aim of the role-playing game is to make the students experience doing a multi-stakeholder assessment in a real research scenario. Each student will play a role of one of the stakeholders involved in the DPIA process. This experience will help the students to apply the concepts learned in the online training or in the class lecture. The game will help the students in taking into account the perspectives of the other people involved in the process.

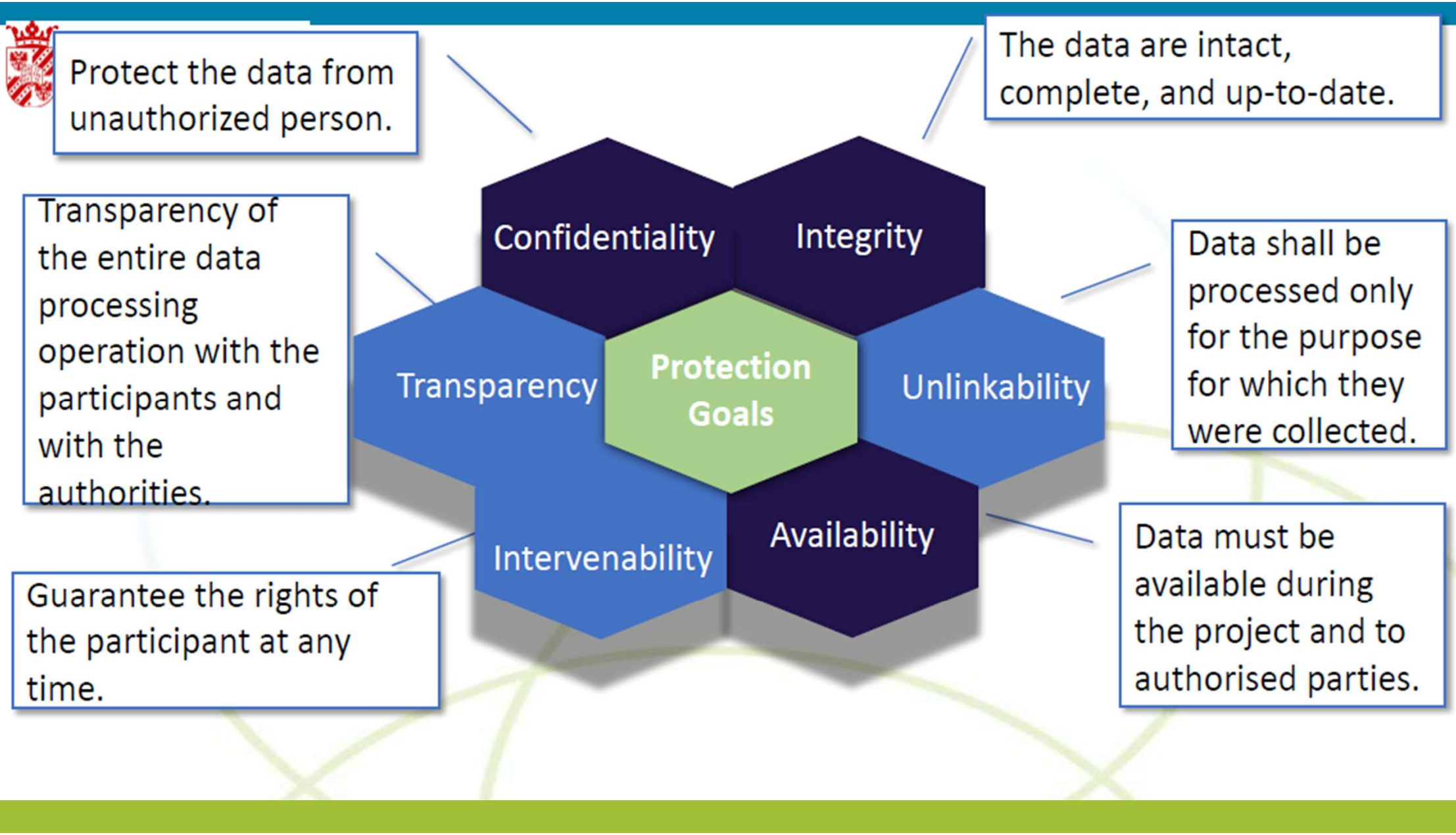
https://sites.google.com/rug.nl/privacy-in-research/role-playing-game

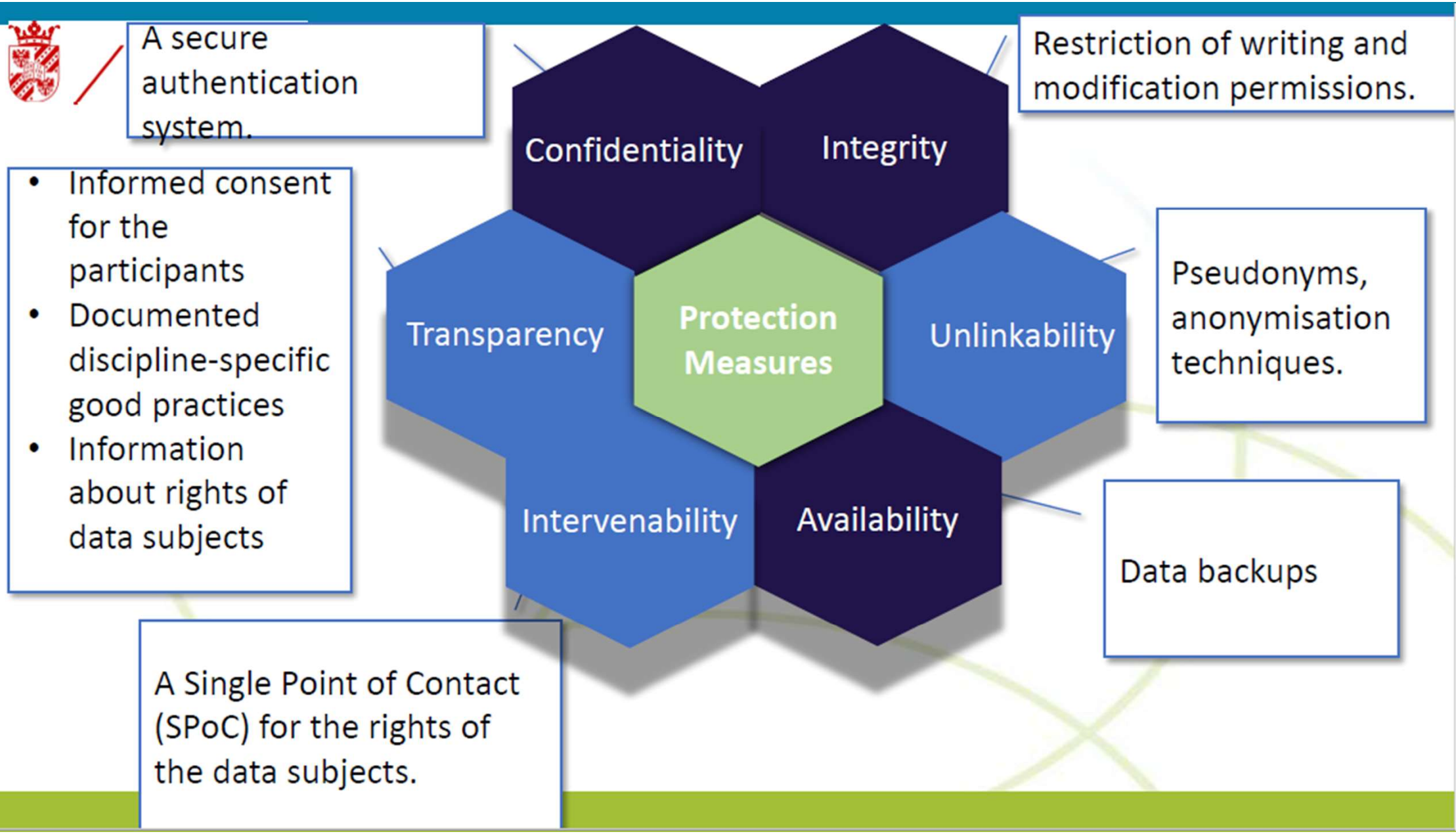
DELAD DPIA Roleplay material

The DPIA Method

It is a **multi-stakeholders** approach that provides a **structured** way of thinking about risks and protection measures.

The risk assessment is based on **protection goals**.


Protect the data from unauthorized person.
The data are intact, complete, and up-to-date.
Transparency of the entire data processing operation with the participants and with the authorities.
Confidentiality
Integrity
Transparency
Protection Goals
Unlinkability
Intervenability
Availability
Data shall be processed only for the purpose for which they were collected.
Guarantee the rights of the participant at any time.
Data must be available during the project and to authorised parties.


A secure authentication system.
Restriction of writing and modification permissions.
Confidentiality
Integrity
Transparency
Protection Measures
Unlinkability
Intervenability
Availability
Informed consent for the participants
Documented discipline-specific good practices
Information about rights of data subjects
Pseudonyms, anonymisation techniques.
Data backups
A Single Point of Contact (SPoC) for the rights of the data subjects.

DELAD DPIA Roleplay material

Our follow up: DELAD the movie !

Use case:

- *Alice, researcher, has developed an algorithm for voice conversion. This technique can alter the speech signal in a way so that the speakers could not be recognised by the third party. This kind of pseudonymised speech data is potentially useful for research in many other disciplines or research design. The algorithm was developed based on speech samples collected from typical adult speakers of English, German, and Dutch. The next step is to test this algorithm on other languages, age groups and speakers with different types of speech disorders. There is a dataset that comprises speech samples of 60+ Polish-speaking children with speech difficulties associated with hearing impairment. The dataset is available under restricted access conditions. The repository sees it as its responsibility to contact the representatives of data providers for permission.*

The video shows a meeting of stakeholders of the Polish dataset, based on a method for a Data Protection Impact Assessment. They discuss if and under which conditions the dataset can be made available for the research purpose in the light of the GDPR.

DELAD DPIA Roleplay material

DELAD the movie ! Recorded in September 2021

Actors in the role play:

- Alice Lee as Alice (postdoc)
- Esther Hoorn as Erin (Legal expert)
- Nicola Bessell as Aine (Ethics Committee member)
- Satu Salaasti as Anita (Researcher)
- Katarzyna Klessa as Kate (Data subject representative)
- Henk van den Heuvel as Archie (Data archiver)
- Paul Trilsbeek as Alan (ICT specialist of the archive)

Material:

- Trailer of the role play
- Video of the role play
- Role cards for the play
- DPIA report created as a result of the role play (with role play introduction)

DELAD DPIA Roleplay material

- The material was presented at the CLARIN Annual Conference 2021.
  - The link will lead you to a description of the educational material
  - The material is used in the master programme on Voice Technology, University of Groningen
  - Interesting for your courses?

- There is also a CLARIN Impact Story about the role play: Navigating the GDPR with Innovative Educational Materials

Recent developments at ACE / DELAD for support of sharing CDS

1. Introducing DELAD
2. CDS annotation tools and techniques
3. Guidelines for consent & storage
4. DELAD DPIA Roleplay material
5. A walk through the DELAD website

DELAD CLARIN