

# Archiving corpora of speech with communication disorders

DELAD workshop 27-28 January 2021

Paul Trilsbeek  
The Language Archive  
Max Planck Institute for Psycholinguistics

# Outline

- Brief introduction of The Language Archive
- General archiving and long-term preservation principles
- CLARIN Centres
- CMDI metadata
- Specific considerations for sensitive data, including CSD
- Accessing and depositing collections at The Language Archive
- Time for discussion/questions

# The Language Archive

- Digital archive of language materials based at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
- Archive exists since the late 90's, initially archiving language materials from MPI's own field researchers and language acquisition researchers
- Became the central archive for the DOBES endangered languages documentation programme, funded by the Volkswagen Foundation in 2000
- 64 TLA collections added to the UNESCO Memory of the World register in 2015
- Research data repository and archive of cultural heritage at the same time
- CLARIN B-Type centre, holds CoreTrustSeal certification

# Collections in The Language Archive

- Provide a unique record of how people around the world use language in everyday life
- More than 350 collections covering more than 250 different languages:
  - Languages from around the world studied by MPI Language and Cognition field linguists
  - First and second language acquisition corpora
  - Endangered languages documented for the VolkswagenStiftung DOBES programme
  - Spoken Dutch corpus
  - Sign language corpora
- More than 15.000 hours of audio and video recordings
- More than 1 million files, about 110 TB of data

# ELAN Software

- Software for creating time-aligned, multi-layered annotations of audio and video files
- Developed by the MPI for over 20 years now
- Widely used by linguists from different sub-disciplines but also by researchers from other domains who have a need for transcribing and annotating audio and/or video
- Available as a downloadable application for Windows, Mac and Linux:  
<https://archive.mpi.nl/tla/elan>

# Archiving and long-term preservation

- Archiving and sharing of research data part of normal scientific practice nowadays
- A lot of focus on “open” science, FAIR data (Findable, Accessible, Interoperable, Reusable)
- Truly “open” data in the sense of unrestricted access obviously not possible for sensitive materials
- Lots of research data repositories exist (2620 listed on re3data.org)
- Generalist vs. Domain repositories: if possible, deposit with a certified\* domain repository as it can better cater for the needs of the particular research community.

\* According to CoreTrustSeal, Nestor Seal / DIN 31644 or ISO 16363

# Archiving and long-term preservation

- Digitisation of analogue media (tape, film, etc.) before they deteriorate
- Digital preservation:
  - Preservation of “bit-streams”: making sure that stored digital objects (files) on a storage medium are preserved -> backups, replication to other locations, timely migration of storage media
  - Preservation of content: making sure that the content of the objects remains interpretable over time -> file format migration before they become obsolete (ideally without loss of information), use of file formats that are well suited for preservation (preferably open standards)
- Good quality Metadata is essential for current and future use of the materials
- Further aspects of “Trusted Digital Repositories”: 16 requirements of the CoreTrustSeal [www.coretrustseal.org](http://www.coretrustseal.org)

# CLARIN “B-type” Centres

- Centres within the CLARIN infrastructure that offer the scientific community access to resources, services and knowledge on a sustainable basis
- Currently 20 certified B-Centres, with different foci, e.g. regional focus or focus on a certain linguistic sub-discipline
- Many of them have a repository for archiving and sharing corpora





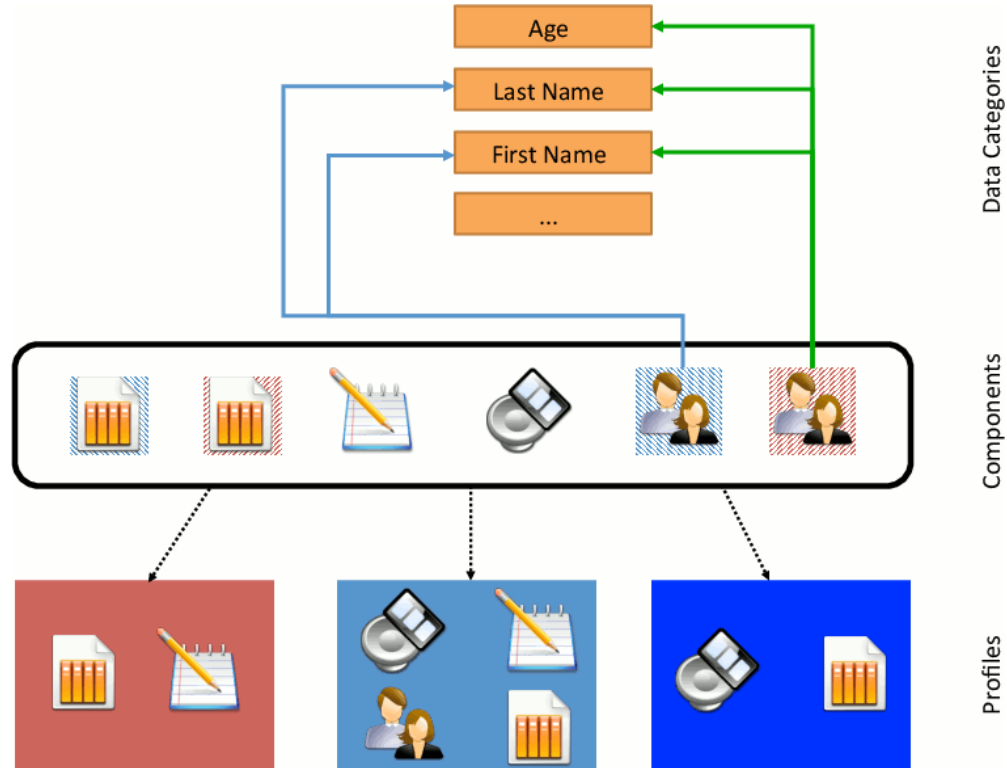
# CLARIN B-Centre Requirements

- Support CLARIN CMDI metadata
- Offer metadata via the OAI-PMH protocol for metadata harvesting
- Support for Shibboleth/SAML2 authentication in order to be part of CLARIN Service Provider Federation
- Support for Persistent Identifiers for the data (e.g. DOI or Handle)
- Repository must meet CoreTrustSeal requirements

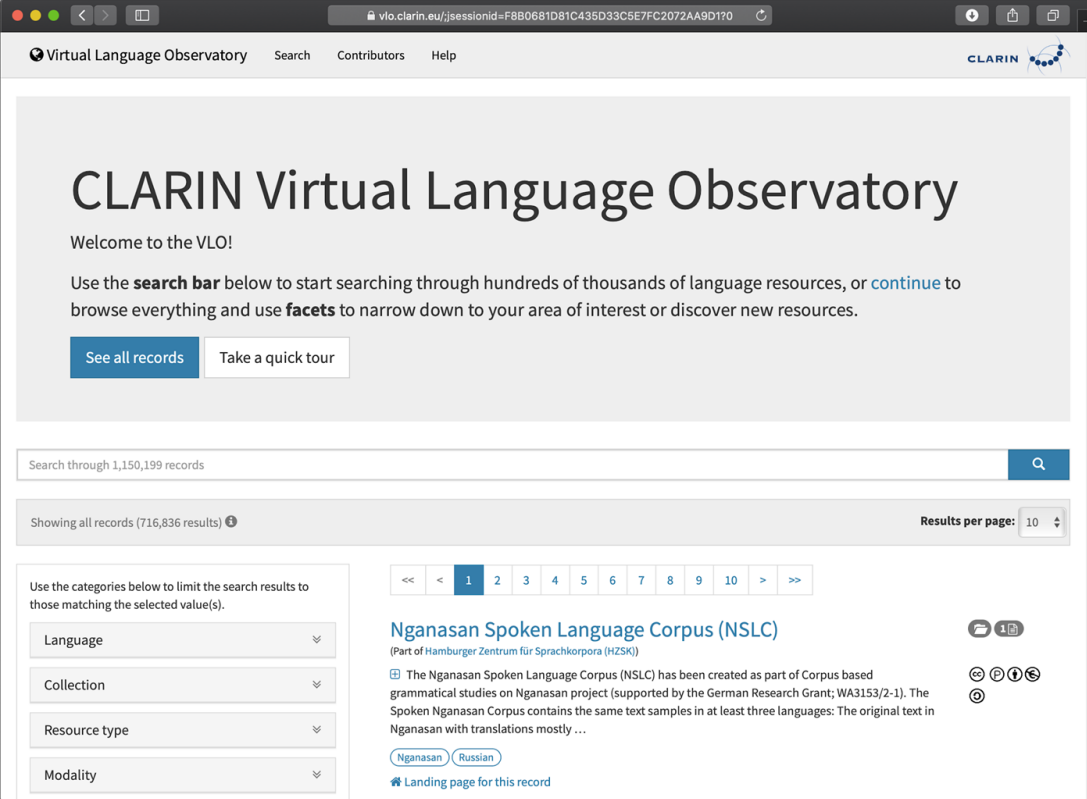
# CMDI metadata

- Component MetaData Infrastructure, developed within CLARIN
- Not a single schema for metadata, but a framework that enables the composition of metadata “profiles” from building blocks that are stored in a central registry
- Enables the creation of metadata profiles that have been tailored for specific types of data, while still ensuring a level of interoperability
- Many CMDI profiles have been created. All are stored in the CLARIN CMDI Component Registry.
- Most CLARIN Centres with a repository require the use of specific CMDI profiles -> Familiarise yourself with the metadata requirements of the repository you will be working with (not just in the case of CLARIN)

# CMDI Metadata: data categories, components and profiles



# CMDI metadata: CLARIN Virtual Language Observatory



The screenshot shows the CLARIN Virtual Language Observatory (VLO) website. The browser address bar displays the URL `vlo.clarin.eu/?sessionId=F8B0681D81C435D33C5E7FC2072AA9D170`. The page header includes navigation links for "Virtual Language Observatory", "Search", "Contributors", and "Help", along with the CLARIN logo.

## CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or [continue](#) to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

[See all records](#) [Take a quick tour](#)

Search through 1,150,199 records

Showing all records (716,836 results)

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Resource type
- Modality

Navigation: << < 1 2 3 4 5 6 7 8 9 10 > >>

### Nganasan Spoken Language Corpus (NSLC)

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

The Nganasan Spoken Language Corpus (NSLC) has been created as part of Corpus based grammatical studies on Nganasan project (supported by the German Research Grant; WA3153/2-1). The Spoken Nganasan Corpus contains the same text samples in at least three languages: The original text in Nganasan with translations mostly ...

[Nganasan](#) [Russian](#)

[Landing page for this record](#)

1/1

CC BY-NC-SA

## Archiving sensitive data

- “Special category” data according to GDPR article 9: data about race, religion, sexual orientation, political option etc. but also genetic data and health data.
- Can only be collected and processed in specific cases. Scientific research is such a case, but extra care needed to ensure that the data will not be misused
- Data Protection Impact Assessment (DPIA) can be helpful to identify risks, define measures to minimise them and to clarify responsibilities (more about this tomorrow in Esther Hoorn’s introduction and the DPIA role play)

## Expected from depositors of TLA

- Anonymising/pseudonymising data as much as possible without rendering it useless for research purposes. I.e. do not include names and other directly identifying information in metadata and textual resources, but do not make voices or faces unrecognisable. Certain “personal” metadata is essential for certain types of research, e.g. age of a child for language development research, or place of birth / current place of residence for dialectology.
- Consent from participants in which they agree with archiving of the data and making it available to users of the archive. In case of speech with disorders, this typically means to other researchers that have received explicit permission. Include a blank consent form with the dataset.

## Technical measures at TLA

- “Data protection by design and by default”:
  - Up-to-date systems and software
  - Secure transport of data (HTTPS)
  - Elaborate system of access policies and authorisation
- At least 6 copies of the archived data are automatically created in different locations to enable disaster recovery
- All archived copies reside within the EU (Netherlands, Germany) at trusted data centres within the Max Planck Society
- Access to systems that hold the data is limited to necessary staff only (system administrators, archive management)

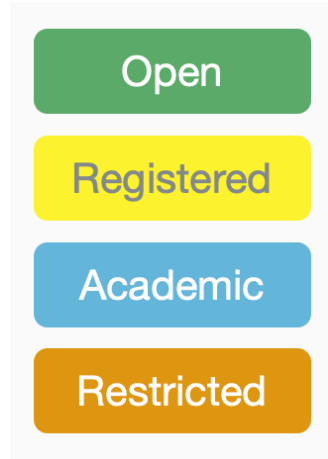
## Technical measures at TLA

- Encryption: No strict requirement in the GDPR but mentioned as an appropriate protection measure
  - Conflict: significant risk with respect to long-term preservation, reduces possibilities for working with data in the archive (e.g. playback of annotated media in web-based viewer)
  - Audio-visual data in TLA is not encrypted “at rest” at the moment
  - Working on a solution that allows us to use encryption while mitigating the issues above
- TLA is CoreTrustSeal certified, which demonstrates compliance with technical and organisational requirements to ensure proper handling and preservation of data. <https://www.coretrustseal.org/wp-content/uploads/2019/01/The-Language-Archive.pdf>



# Access policies

- Access levels:
  - Open: Completely public, no registration required
  - Registered: Accessible to any registered user
  - Academic: Accessible to any academic user
  - Restricted: Access permission needs to be requested on an individual basis, depositor typically decides
- Depositor determines appropriate access level
- Different access policies can be defined for different files within a collection, if necessary
  - E.g. anonymous transcriptions could be “open” and audio recordings “restricted”



# Authorisation

- Account registration: validity of a new user is manually verified by archive staff, optionally their academic status as well
- Login via own academic account (“Shibboleth” federated login): account is active immediately and academic status is automatically assigned
- Access to specific restricted materials is granted to individual users once a request is approved
- Access request: Users needs to specify intended usage of the resources

# Licensing/Agreements

- Agreements needed for archiving and sharing:
  - Deposit / Processing agreement, between depositor and archive
  - Data use agreement / License / Terms of Use, between archive and data user
- Both types currently under review to investigate whether changes are needed in terms of GDPR-compliance. E.g. most existing end user license agreements that are used for language data are “perpetual”, which seems to be in conflict with the principle of keeping data only as long as is necessary
- CLARIN "RES" (restricted use) licenses most appropriate, but also do not address the duration of use aspect yet:

<https://www.clarin.eu/content/licenses-and-clarin-categories>

## TalkBank exchange

- Arrangement with TalkBank at Carnegie Mellon University in Pittsburgh, U.S.A. for sharing non-sensitive (anonymous/anonymised) parts of collections of “atypical” speech
- Anonymous metadata and transcriptions/annotations are made publicly available in the appropriate part of the TalkBank system
- Both non-sensitive and sensitive files (audio, video) are stored at The Language Archive
- Links to the collections at TLA are provided on the TalkBank collection overview pages

# General GDPR considerations

- As with any complex law, there's room for different interpretations. No case law yet for the GDPR in relation to academic data
- Different local implementations make things even more complex in an international setting
- Different approaches:
  - Better safe than sorry: stop sharing. Clearly in conflict with current views on proper scientific conduct ("FAIR" principles)
  - Comply to the best of our ability, knowing that there's a chance that we may not always be 100% compliant, in particular with older pre-GDPR collections
- The incentive for the EC to create the GDPR was not to make it hard or impossible for the scientific community to collect, share and preserve personal data for research purposes

# The Language Archive: [archive.mpi.nl](http://archive.mpi.nl)

The screenshot shows a web browser window with the URL [archive.mpi.nl](http://archive.mpi.nl). The page features the Max Planck Institute for Psycholinguistics logo at the top left. Below the logo, there are two main sections:

**The Language Archive**: This section includes a world map with numerous location markers and a text block stating: "The Language Archive at the Max Planck Institute in Nijmegen provides a unique record of how people around the world use language in everyday life. It focuses on collecting spoken and signed language materials in audio and video form along with transcriptions, analyses, annotations and other types of relevant material (e.g. photos, accompanying notes)."

**MPI for Psycholinguistics Archive**: This section features a photograph of a modern building with large glass windows and a text block stating: "The MPI for Psycholinguistics Archive contains research data from the various (current and former) departments and research groups within the Max Planck Institute for Psycholinguistics in Nijmegen. It includes data resulting from neurobiological studies, genetics studies, and behavioural studies."

The footer of the page contains contact information for the Max Planck Institute for Psycholinguistics, along with logos for the United Nations Educational, Scientific and Cultural Organization (UNESCO), the Core Trust Seal, CLARIN B Centre, and ICSU World Data System.

## Sources for help (in addition to DELAD)

- CLARIN ACE Knowledge Center: <https://ace.ruhosting.nl>
- CLARIN national helpdesks: <https://www.clarin.eu/content/support>
- CLARIN Legal Issues Committee (CLIC):  
<https://www.clarin.eu/governance/legal-issues-committee>
- Staff of the data repository you will be working with
- Local Research Data Management support
- Local Data Protection Officer

# Questions?



Paul.Trilsbeek@mpi.n

| archive.mpi.nl