NELF

A Data Management Plan created using DMPonline.be

Creator: Hugo Van hamme

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Hugo Van hamme, First Name Surname

Data Manager: First Name Surname

Project Administrator: First Name Surname

Grant number / **URL:** S004923N

ID: 197048

Start date: 01-10-2022

End date: 30-09-2026

Project abstract:

Automatische spraakherkenning (speech-to-text) heeft recent grote stappen vooruit gezet, vooral als het om het Engels gaat. In het Vlaams, loopt de technologie echter achter. Dat komt voornamelijk door een gebrek aan gelabelde data. Met "gelabelde data" wordt bedoeld: opnames van spraak met een letterlijke (verbatim) transcriptie. Dit soort data is beschikbaar in het Engels, maar niet in het Vlaams. De aanmaak van dergelijke data is erg duur: makkelijk 5 arbeidsuren per uur aan spraak.
In NeLF proberen we dit euvel te verhelpen op tween manieren: (1) de hoeveelheid data verhogen (2) inzetten op ongesuperviseerd of zwak-gesuperviseerd leren. Het voordeel van de tweede techniek is dat goedkope data ingezet kan worden.
Met zwak-gesuperviseerd leren wordt bedoeld dat modellen voor spraakherkenning geleeerd worden adhv spraakopnames vergezeld van een niet-verbatim transcript, bijvoorbeeld ondertitels in televisieuitzendingen of een verslag van een vergadering.
Met ongesuperviseerd leren wordt bedoeld dat spraakmodellen gelerd worden uit enkel de spraak, zonder transcript.

Last modified: 10-03-2023

NELF
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

The type of data used in NeLF are audio recordings of people speaking in various situations including monologues, discussions and read speech.
The audio is accompanied by metadata which could include a verbatim transcription, a summary, a report. When available, the identity and/or their age and/or their gender of the speaker(s) in the audio documents may be stored.
The data will mostly be a copy of data recorder by third parties. Data providers will be VRT broadcasts, recordings of parliament and city council meetings, ...
We will also setup a data donation website, where companies and citizens can submit their data after signing an informed consent form.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. For legal issues Toon.Boon@kuleuven.be
2. The data is an estimated 4 TB in size. It will consist of a corpus accessible to the university parties only and a public corpus. Both will be curated at KULeuven's RDR long-term preservation service (https://www.kuleuven.be/rdm/en/rdr) at a cost of 113 euro/TB/year. The public corpus will be made available for download via the universities' websites and via Instituut voor Nederlandse Taal, which has a data curation and distribution service as well.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

None, unless the agreement with the data provider does not allow us to do so.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Speech data is (non-sensitive) personal information, since people are (to some extent) recognisable by their voice.
Our data provider partners do not want part of the data to be made public. Therefore, we have segregated the two corpora mentioned above.
The private corpus will be encrypted. The key for decryption will reside on a key server, so illegal copies are void without the key.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

none

NELF
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

Question not answered.

NELF
GDPR

GDPR

Have you registered personal data processing activities for this project?

Question not answered.

NELF
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Subtitled media B1	TV programs with subtitles	reuse existing data	digital	observational	audio with aligned text .wav, .xml	<10TB	
Public meetings B2	City councils, parliament meetings	reuse existing data	digital	observational	audio with textual meeting reports .wav, .xml	<5TB	
B3	Private meetings, seminars, lectures	generate new data	digital	observational	audio with aligned text .wav, .xml	<50 GB	
B4	Crowd sourcing	generate new data	digital	observational	audio with aligned text .wav, .xml	<50 GB	
Recording list	For each recording (when available): list of pseudonymized speakers occurring in each recording with speaking time, year of recording of the audio.	generate new data	digital	compiled	.xlsx	<100 MB	
Speaker list	For each pseudonumized speaker, age, gender and postal code where raised. Data fields may be missing.	generate new data	digital	compiled	.xlsx	<100 MB	
ASR models	Speech models suitable for automatic speech recognition in which contributions of participants are not recognizable.	generate new data	digital	compiled	.json, .ep	<100 GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Data from private archives of media companies such as VRT.
City council meeting recordings and recordings of Flemish Parliament, e.g. https://www.vlaamsparlement.be/nl/parlementair-werk/de-plenaire-vergadering

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

The human data are voice recordings, which are personal data.
However, we see no ethical issues and have covered privacy issues:
1) in the project, our goal is to build speech models for automatic speech recognition. From these models, the contribution of each participant cannot be recovered (inverted).
2) Any data that will be made available to other researchers will have the license restriction it can only be used to build models from which the contributions of individuals is cannot be recovered.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

Voice recordings, pseudonym, age, gender, postal code of place where person was raised (as a description of regiolect)

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Data transfer agreement for B1 with VRT signed. Corpus will be extended throughout the project under similar data transfer agreements.
DTA stipulates data cannot be made public.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

For B1-B4 jointly, there will be two spreadsheets where we list
1. for each recording the duration of speech for each speaker occurring in the recording
2. for each speaker: pseudonymized identity, age, gender and postcode where the person grew up.
This will be used to steer data selection efforts and document B1-B4.
See also section 1.1.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

see previous point

3. Data storage & back-up during the research project

Where will the data be stored?

On the servers with RAID disks of KU Leuven and U Gent.
User access restrictions are implemented.

How will the data be backed up?

B1 and B2 are backed-up on tape drive. Notice also that risk of data loss is limited: audio recordings are a copy from archives (e.g. at VRT). Secondly, there will be a copy at KU Leuven and at U Gent.
B3 and B3 are smaller is size and have daily incremental backups at both universities. There will be a copy at KU Leuven and at U Gent.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

150 TB RAID disk.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Linux Access Control Lists can define access rights for each user/file combination.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Backup device and tapes purchased from project funds. About 8000 euro.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Corpora B1 and B2 are subject to Data Transfer Agreements. For instance, our agreement with VRT for B1 stipulates data copies need to be destroyed 2 years after the end of the project. This is very likely to be the case for B2 as well. (Data Protection Officers have other objectives than reproducible research.)

Where will these data be archived (stored and curated for the long-term)?

Corpora B1 and B2 and parts of speaker and recording list pertaining to B1 and B2 will remain on our servers for 2 years after project end.
Corpora B3 and B4 and parts of speacer and recording list pertaining to B3 and B4 will be made available through Instituut voor Nederlandse Taal (https://ivdnt.org/taalmaterialen/)
ASR models will be made available through huggingface.co, where itcan be found easily.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

None

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in a restricted access repository (after approval, institutional access only, …)
- No (closed access)

no for B1 and B2 and related parts of speaker and recording list.
yes for B3 and B4 and related parts of speaker and recording list.
ASR models will be made available.

If access is restricted, please specify who will be able to access the data and under what conditions.

For B1 and B2, only the project collaborators will have access.
For B3 and B4 anyone who signs the data agreement with Instituut voor Nederlandse Taal.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects

B1: data transfer agreement with VRT restricts access to KU Leuven and U Gent
B2: no agreements made yet, but same restrictions are likely to apply
B3 and B4: participants will be informed data will be shared.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

B1 and B2: NA
B3 and B4: Instituut voor Nederlandse Taal
ASR models: huggingface.co

When will the data be made available?

Prior to 30 Sept 2026 (project end).

Which data usage licenses are you going to provide? If none, please explain why.

License will not allow for re-sharing.
Each new user needs to sign the agreement with Instituut voor Nederlandse Taal.
Agreements stipulate that data can be used to build new products (e.g. speech models) in which the voice of participants in corpora B3 and B4 is not recognizable.
This protects the privacy of participants and excludes deep fake applications.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

None

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Hugo Van hamme

Who will manage data storage and backup during the research project?

Hugo Van hamme

Who will manage data preservation and sharing?

Hugo Van hamme

Who will update and implement this DMP?

Hugo Van hamme