

---

## SPASS

*A Data Management Plan created using DMPonline.be*

**Creator:** Hugo Van hamme

**Affiliation:** KU Leuven (KUL)

**Template:** KU Leuven BOF-IOF

**ID:** 197935

**Last modified:** 21-03-2023

# SPASS

## 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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: <b>N</b> (ew data) or <b>E</b> (xisting data)	Indicate: <b>D</b> (igital) or <b>P</b> (hysical)	Indicate: <b>A</b> udiovisual <b>I</b> mages <b>S</b> ound <b>N</b> umerical <b>T</b> extual <b>M</b> odel <b>S</b> oftware <b>O</b> ther (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
Speech recordings	sound recordings of pupils performing a reading task	N	D	S	.wav	<100GB	
Annotation	for each sound file, what should have been read, what was read, miscues	N	D	T	.xml	<1GB	
Metadata	for each participant: age, sex, language spoken at home, AVI level, sound file names, annotation file names	N	D	T	.xlsx	<1GB	

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:

no

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- Yes, human subject data (Provide SMEC or EC approval number below)

Approval pending  
PRET [G-2023-6272](#)

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- Yes (Provide PRET G-number or EC S-number below)

PRET [G-2023-6272](#)

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

Societal.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or 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.

- No

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

## 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).

Every sound file is annotated by master-level students.

See data types.

We follow the same annotation protocol as used in our earlier work: <https://taalmaterialen.ivdnt.org/download/tstc-chorec-spraakcorpus/>

**Will a metadata standard be used to make it easier to find and reuse the data ?**

**If so, please specify which metadata standard will be used.**

**If not, please specify which metadata will be created to make the data easier to find and reuse.**

- No

meta data provides relevant speaker information. See data types.

## Data Storage & Back-up during the Research Project

**Where will the data be stored?**

- Other (specify below)

Unix data servers at ESAT (dept. electrical engineering).

These implement access control lists and have a daily backup.

**How will the data be backed up?**

- Other (specify below)

Backups by ESAT system managers.

**Is there currently sufficient storage & backup capacity during the project?**

**If no or insufficient storage or backup capacities are available, explain how this will be taken care of.**

- Yes

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

Access control lists specify the privileges per user.

Only project collaborators will have access.

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

No extra cost charged.

## Data Preservation after the end of the Research Project

**Which data will be retained for 10 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...).**

- All data will be preserved for 10 years according to KU Leuven RDM policy

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

- Other (specify below)

At Instituut voor Nederlandse Taal, where the precursor of this data is also available: <https://taalmaterialen.ivdnt.org/download/tstc-chorec-spraakcorpus/>

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

None for us. Covered by INT.

## Data Sharing and Reuse

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

- Yes, as restricted data (upon approval, or institutional access only)

Data will be downloadable from Instituut voor Nederlandse Taal.

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

Users will need to sign a license that restricts the usage of the data. Most importantly, derived products cannot contain the participant's data in a recognizable form. This allows to e.g. build models for automatic recognition of speech, but not for text-to-speech applications.

**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 per dataset or data type where appropriate.**

- Yes, privacy aspects

to the extent that persons are recognizable from their voice.

But then again, WHAT they say is not sensitive information but AVI-standardized texts (stories, word lists).

**Where will the data be made available?**

**If already known, please provide a repository per dataset or data type.**

- Other data repository (specify below)

Instituut voor Nederlandse Taal.

**When will the data be made available?**

- Specific date (specify below)

no later than 30 Sept 2026

**Which data usage licenses are you going to provide?**

**If none, please explain why.**

- Other (specify below)

Users will need to sign a license that restricts the usage of the data. Most importantly, derived products cannot contain the participant's data in a recognizable form. This allows to e.g. build models for automatic recognition of speech, but not for text-to-speech applications.

Examples of this agreement are available for the corpora downloadable from INT now. E.g. license of <https://taalmaterialen.ivdnt.org/download/tstc-couc/>

**Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.**

- No

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

none

## Responsibilities

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

Gloria Montoya, supervised by Hugo Van hamme and Pol Ghesquiere.

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

ESAT system managers supervised by Hugo Van hamme

**Who will manage data preservation and sharing?**

Hugo Van hamme will make sure the data are available at INT (or find an alternative).

**Who will update and implement this DMP?**

Hugo Van hamme