Identifying Dyslexia Earlier in Life
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)

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
CTSiN The role of reading experience in atypical cortical tracking of speech and speech-in-noise in dyslexia	Open source available dataset which contains MEG data of children with and without dyslexia listening to speech (presented with or without lip movements / with or without noise)	*E (existing)*	**D (digital)**	Audiovisual Sound **N**umerical Textual	.fif	<300GB	
DYSCO (-HONEY)	We plan to *collect* a dataset, collecting neural data (EEG and MRI), speech recordings and behavioural data as part of a longitudinal study that will follow preschool children across development. A more extensive description is added below.	***New***	D (digital) & P (physical)	Audiovisual Sound **N**umerical Textual	.pdf . apr .txt .csv .pdf EEG: .bdf MRI: .bval, .bvec, .par, .rec, .xml, .json, .nii	~ 7 TB	Physical data (e.g. outcomes from paper pencil tests). This data does not contain participant information and will be stored in closed cabinets (10th floor ON2, ExpORL).

Important note regarding the DYSCO-HONEY dataset: This dataset will be collected in collaboration with several PhD and Postdoctoral researchers. For the current project, I will focus specifically on a subset of the collected data, namely the behavioral and EEG data. As a result, this Data Management Plan (DMP) will concentrate on these file types. Data curation for the full DYSCO-HONEY dataset is outlined in a separate DMP (HONEY project: Heterogeneity and Oscillome in Naturalistic Environments in Young Children; ID 212074).

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. Designation of responsible person (If already designated, please fill in his/her name.)
   - CTSiN dataset: The source data is online available (https://osf.io/9ce5t/). We are not responsibility of maintaining this source data, no responsible person has been designated on this behave.
   - DYSCO dataset: prof. Maaike Vandermosten will be responsible for preserving the data at least 5 years after the end of the research.
2. Storage capacity/repository
   - during the research
     - CTSiN dataset: On personal, encrypted hard drive.
     - DYSCO dataset: On personal, encrypted hard drive and on KU Leuven maintained drives (see more details in DMP 212074)
   - after the research
     - CTSiN dataset: A compressed format of the data will be stored on the KU Leuven maintained drives together with

    - the intermediate results and processing code if the dataset is used for publication. This compressed format should enable replication of published study.
    - DYSCO dataset: The source data will be stored on KU Leuven maintained drives. Note that one data repository will be created for the whole dataset containing all the gather files from behavioural, EEG and MRI measurements. (see more details in DMP 212074)

Remark on Storage Capacity During Research:

Given the large size of source data files (e.g., raw EEG data), it is necessary to use a local hard drive for efficient processing which is not guaranteed when storing these large files on network drives. However, it is essential that the local hard drive contains duplicate copies of the files stored on the network drives. All source data must remain available on the network drives at all times.

In addition to the raw data, processing code will also be generated throughout the research. This code will be stored in both GitLab and GitHub repositories, with a Git tag created upon publication. This practice ensures that the code archived is an exact, reproducible version of the one used in the publication.

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)

NA

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)

In both datasets, the neural responses of children were/will be collected to speech. Children are a vulnerable population and therefore specific security measures should be taken.

- CTSiN dataset: The raw data (MEG responses and behavioural data) is available online. However, no personal data which allows direct identification of the participant is shared. Therefore no specific security measures are taken.
- DYSCO dataset: The data set maintained on the KU Leuven maintained drives combines the information for behavioural, EEG and MRI experiments. However, the information linked to identification of the participants (obtained in files such as participant questionnaire and informed consent) will not be uploaded in this dataset. These files, linked to the identify of the participant, will be stored in a separate folder, with limited access only accessible for the researchers involved in the data collection.

The copy on the personal hard drive shall never contain files to allow direct identification of the participant. An additional security measure is taken by encrypting (i.e., password protecting) the hard drive.

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

We will implement two key steps to enhance the readability of the datasets and the reproducibility of the results:

1. **Implementing the BIDS Format:** The BIDS (Brain Imaging Data Structure) format will be used to organize the data, which involves incorporating additional metadata files. This structure will improve both the readability and reproducibility of the dataset, as well as the derived results. Metadata will be added whenever available.
2. **Code Maintenance Using GIT (GitLab):** GitLab will be utilized for managing and versioning the code. It will also serve as a backup for all scripts and code necessary to generate intermediate results, ensuring that the entire process is reproducible and well-documented.

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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Digital Data Type	File format	Data volume	Physical volume
CTSiN The role of reading experience in atypical cortical tracking of speech and speech-in-noise in dyslexia	Open source available dataset which contains MEG data of children with and without dyslexia listening to speech (presented with or without lip movements / with or without noise)	*E (existing)*	**D (digital)**	**A**udiovisual **S**ound **N**umerical **T**extual Observational Software (code)	.fif	<300GB	
DYSCO-HONEY	We plan to *collect* a dataset, collecting neural data (EEG and MRI), speech recordings and behavioural data as part of a longitudinal study that will follow preschool children across development. A more extensive description is added below.	***New***	D (digital) & P (physical)	**A**udiovisual **S**ound **N**umerical **T**extual	.pdf . apr .txt .csv .pdf EEG: .bdf MRI: .bval, .bvec, .par, .rec, .xml, .json, .nii	~ 7 TB	Physical data (e.g. outcomes from paper pencil tests). This data does not contain participant information and will be stored in closed cabinets (10th floor ON2, ExpORL).

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:

The dataset (CTSiN) is available through this link: https://osf.io/9ce5t/

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

In both datasets, the neural responses of children were/will be collected to speech. Children are a vulnerable population and therefore ethics should be considered.
Ethical approval:

- CTSiN dataset: The study which collected the dataset was approved by the local ethics committee (Comité d’Ethique Hospitalo-Facultaire Erasme-ULB, 021/406, Brussels, Belgium; approval number: P2017/081). Since this project only will

reuse the existing data from the public available dataset, no ethical approval by EC is required (confirmed by EC Leuven on 20/12/2023).
- DYSCO dataset: Ethical approval (S70080) is currently under review.

The one of the largest issues, after collection of the neural data, is protection of the child's identify. We will take all necessary measures to make sure no direct identification is possible from the neural data, i.e., no files with identity details will be stored in the overall dataset, the files with identify details will be stored in a separate folder on KU Leuven maintained storage, with limited access only accessible for the researchers involved in the data collection.

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

Personal data will be collected, purely for practical purposes (i.e., contacting the participant for collection of later time point, deposit of compensation etc). However, except for practical purposes, this personal data will not be processed for further research purposes.
The collected data (i.e., behavioural, EEG and MRI outcomes) will be pseudonymised before storing the data on KU Leuven maintained drives. The key for this pseudonymisation will be stored in a separate folder on KU Leuven maintained storage, with limited access only accessible for the researchers involved in the data collection.

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.

- 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

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).

We will implement two key steps to enhance the readability of the datasets and the reproducibility of the results:

1. **Implementing the BIDS Format:** The BIDS (Brain Imaging Data Structure) format will be used to organize the data, which involves incorporating additional metadata files. This structure will improve both the readability and reproducibility of the dataset, as well as the derived results. Metadata will be added whenever available.
2. **Code Maintenance Using GIT (GitLab & GitHub):** GitLab & GitHub will be utilized for managing and versioning the code. It will also serve as a backup for all scripts and code necessary to generate intermediate results, ensuring that the entire process

is reproducible and well-documented.

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.

- Yes

The BIDS data structure incorporates additional metadata files to improve the readability and reproducibility of the dataset. These metadata will be included whenever available. BIDS is a standardized format that ensures proper documentation of data with sufficient metadata for easy readability.

At ExpORL, we use a customized variation of the BIDS structure to support additional file types. However, this custom variation still maintains the mandatory metadata requirements of the standard BIDS format.

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

Where will the data be stored?

- CTSiN dataset: The raw data is stored online (https://osf.io/9ce5t/). In order to analyze the data, some files will need to be stored on the (encrypted) hard drive (since calculations from a non-local source are too slow and lead to computational failures).
- DYSCO dataset: The source data will be stored on KU Leuven managed drives. A local copy is save on the (encrypted) hard drive. Note that the duplicate files on the encrypted hard drive will never be files which can be linked to identification of the participant.

The generated script will be stored on Gitlab/GitHub.

How will the data be backed up?

A backup on the encrypted local hard drive is taken automatically at least once a week.

GitLab/GitHub manages the code and its versions, and therefore, also functions as a back up of all code to generate intermediate results.

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

The local personal encrypted hard drive has a capacity of 4 TB. If additional storage is needed, an external hard drive can be purchased. To ensure sufficient backup space, an external 10 TB hard drive has already been acquired.

Furthermore, the data will also be stored on the L and K drives, which are backed up according to KUL's backup strategies. The partitions on both the L and K drives will be incrementally expanded as necessary to accommodate the growing dataset over time.

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

The data on the external hard drives is protected through encryption, ensuring that access is restricted. The encryption is password-protected, allowing only the designated researchers to access their personal hard drives.

On the **L-drive**, permission settings are configured so that files can be modified by researchers involved in the DYSCO collaboration. Once data is moved to the **K-drive**, files are locked and cannot be modified.

The research scripts are securely backed up through **GitLab & GitHub**, which is accessible to all team members and principal investigators (PIs), but restricted from unauthorized individuals.

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

Expected Costs for data storage and backup during the DYSCO HONEY project:

1. **L-drive**:
   €475.70 per year x 4 years = **€1 902.80** (for 5 TB)
2. **K-drive**:
   - For the first 2 years: €4.76 per 100 GB per year x 35 x 2 year = €333.2 (for 3.5 TB)
   - For the following 2 years: €4.76 per 100 GB per year x 70 x 2 year= €666.4 (for 4 TB)
     Total for K-drive: €999.6
3. **Total Storage Cost on KUL Maintained Drives** :
   €2 902.4, which will be covered by the C1 bench fee.
4. **Imaging-Specific Hard Drives**:
   The imaging-specific hard drives, purchased for **€722.99**, were acquired using the personal FWO bench fee allocated to Marlies Gillis.

For more information, see the DMP of the DYSCO-HONEY project (HONEY project: Heterogeneity and Oscillome in Naturalistic Environments in Young Children; ID 212074).

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...).

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)?

- CTSiN dataset: The data is stored online (https://osf.io/9ce5t/). A copy of the data will be stored on KU Leuven maintained drives, when this dataset is used for publication.
- DYSCO dataset: The data will be stored on KU Leuven maintained drives.

The processing code will be stored using git (and a git tag will be generated).
After finalizing the study, the caches and results will be stored on KU Leuven-managed drives (for min. 10 years).

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

Expected Cost of Data Preservation: 10 years x €4.76 per 100 GB per year x 70 = **€3 332** (for 7 TB)
These costs will be covered by grant attributed to prof. Maaike Vandermosten (who is responsible for data preservation).

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.

- Other, please specify:

- CTSiN dataset: This dataset is already open-accessible data. This dataset can be found here: https://osf.io/9ce5t/. If the dataset is used for publication, the processing code to generate the studies findings will be made available.
- DYSCO dataset: Choices regarding data collection are made with having the idea of making the dataset public after finalizing the longitudinal study. However, no details regarding the practicalities of the data sharing has been agreed upon among the PIs guiding this longitudinal project.

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

NA

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

If the DYSCO dataset (at the end of the longitudinal study) is published, the pseudonymised data will be shared to guarantee identification of the participants.

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

Not determined yet.

When will the data be made available?

If the data of the DYSCO dataset will be made available, this will be after finalization of the longitudinal study (i.e., 2028).

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

Not discussed yet.

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.

- Yes

If the data of the DYSCO dataset will be made available, we intend to publish this dataset and therefore this will also obtain a DOI number.

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

Not discussed yet.

6. Responsibilities

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

I (Marlies Gillis) will update and implement this DMP with respect to how I maintain a subset of the DYSCO dataset.

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

I (Marlies Gillis) will update and implement this DMP with respect to how I maintain a subset of the DYSCO dataset.

Who will manage data preservation and sharing?

Not discussed yet.

Who will update and implement this DMP?

I (Marlies Gillis) will update and implement this DMP with respect to how I maintain a subset of the DYSCO dataset.