HONEY project: Heterogeneity and Oscillome in Naturalistic Environments in Young Children

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
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).
DYSCO 2018	We plan to *reuse* a dataset, collected in a previous longitudinal study which consist of structural MRI scans of 330 prereading children.	Reuse	D (digital)	**N**umerical Textual	.pdf . apr .txt .csv .pdf MRI: .bval, .bvec, .par, .rec, .xml, .json, .nii	~ 1.5 TB	

Important Note: During data collection, the physical outcomes generated from paper-and-pencil tasks will be stored in different locations depending on the scientist managing the behavioral task (either Ortho or ExpORL). Once data collection is complete and the behavioral results have been cleaned (after each time point in the longitudinal study), the physical data will be securely transferred to a locked cabinet on the 10th floor of ExpORL. Access to this cabinet will be restricted to researchers involved in the DYSCO collaboration.

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 DYSCO-2018 was collected within the same DYSCO collaboration (collaboration between prof. Jan Wouters, prof.Pol Ghesquière en prof.Maaike Vandermosten). This resulted in the publications listed below. However, the dataset was never made publicly available.

Vanderauwera et al. Early dynamics of white matter deficits in children developing dyslexia. Dev Cogn Neurosci 27, 69–77 (2017).

Blockmans et al. Role of Family Risk and of Pre-Reading Auditory and Neurostructural Measures in Predicting Reading Outcome.

Neurobiology of Language 4, 474–500 (2023).

Economou et al. Investigating the impact of early literacy training on white matter structure in prereaders at risk for dyslexia. Cereb Cortex 32, 4684–4697 (2022).

Vandermosten et al. A DTI tractography study in pre-readers at risk for dyslexia. Dev Cogn Neurosci 14, 8–15 (2015).

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)

This project relies on data collected in children, a vulnerable participant population. An ethical approval for collecting new data is pending.

- Existing dataset (DYSCO-2018): ethical approval was obtained when collecting the dataset ((S55139 and S54415)
- DYSCO-HONEY dataset: ethical approval is under review (S70080)

The one of the largest issues is protection of the child's identity. We will take all necessary measures to make sure no direct identification is possible from the 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, 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)

Yes, G-2025-9003.

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 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).

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.

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.

- 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.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Other (specify below)

DYSCO-2018 Dataset: Currently stored at GBW-0101_DYSCO_Projects2020.

DYSCO Dataset Storage and Management:

The source data for the DYSCO dataset will be stored on KU Leuven-managed drives. During data collection, the data will initially reside on the L-drive, allowing researchers to make updates and modifications as needed. Once the data is converted to the BIDS format, it will be moved to the K-drive, where it will be locked in its final, unmodifiable form.

Researchers will aim to reformat the data to BIDS standards as soon as possible after collecting data from each participant, with the formatted data subsequently transferred to the K-drive for archival. The BIDS-compliant dataset on the L-drive will also contain a "derivatives" folder, where ongoing analyses and intermediate results are stored. Once a study is completed and formatted for publication, the study's code, processing results, and final outputs will also be moved to the derivatives folder on the K-drive. This ensures that all associated data, intermediate findings, and results that lead to the study's conclusions are properly archived.

Imaging Modality Storage:

Each imaging modality (behavioral data, EEG, and MRI) is backed up on a 4TB encrypted hard drive. These drives act as an intermediary to transfer the collected data from the stimulation/recording PC to the L-drive, ensuring redundancy. Each modality's drive contains a copy of the corresponding data on the L-drive.

Data Analysis:

During analysis, data is temporarily stored on the researcher's encrypted hard drive. This is necessary due to the large file sizes, as reading from the network drives would cause significant delays in processing. It is important to note that only pseudoanonymised data is stored on the researcher's local hard drive, ensuring that no identifiable participant information is present.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

The collected DYSCO data is stored and backed up in three different locations:

1. **L-drive**: During ongoing data collection, the data is stored on the L-drive. This drive is regularly backed up according to KU Leuven's backup strategies.
2. **K-drive**: Once the collected data is transformed into the BIDS format, it is transferred to the K-drive. Researchers are encouraged to upload each participant's data to the K-drive as soon as possible after collection. The K-drive is also backed up in accordance with KU Leuven's backup protocols.
3. **Imaging-Specific Hard Drive**: For each imaging modality, the data is first stored on a dedicated encrypted hard drive before being transferred to the L-drive. This imaging-specific hard drive serves as an intermediate storage solution.

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

The size of the required partitions on both the **L-drive** and **K-drive** will be incrementally increased to accommodate the growing dataset over time.

Additionally, the **imaging-specific hard drives** are purchased and encrypted prior to the start of data collection. These drives ensure secure, storage for each imaging modality before the data is transferred to the L and K-drive.

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. For the imaging-specific hard drives, access is limited to researchers involved in the DYSCO collaboration.

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**, 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 research 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.

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

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)?

- Large Volume Storage (longterm for large volumes)

The data will be stored for a longer period using the K-drive.

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).

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 open data

The decisions regarding data collection for the DYSCO-HONEY dataset are made with the intention of eventually making the dataset publicly available after the completion of the longitudinal study. However, the specific details regarding the practicalities of data sharing have not yet been agreed upon among the PIs overseeing the project. It is important to note that any data that could potentially identify participants will not be shared.

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 per dataset or data type where appropriate.

- Yes, privacy aspects

If the DYSCO-HONEY dataset is published at the end of the longitudinal study, only pseudonymized data will be shared to ensure the protection of participants' identities.. The key of this pseudonymization will only be available to the PI after finalization of the project.

Where will the data be made available?

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

- Other (specify below)

Not determined yet.

When will the data be made available?

- Upon publication of research results
- Specific date (specify below)

If the data of the DYSCO-HONEY 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.

- Other (specify below)

Not discussed yet.

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.

- Yes, a PID will be added upon deposit in a data repository

If the data of the DYSCO-HONEY 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.

Responsibilities

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

Each researcher is responsible for creating the proper metadata to their collected data. However, overall for the final dataset, Marlies Gillis, is responsible to ensure all the collected data is correctly combined in one BIDS dataset of the DYSCO-HONEY project.

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

Overall for the final dataset, Marlies Gillis, is responsible to ensure all the collected data is correctly combined in one BIDS dataset of the DYSCO-HONEY project. This dataset will be stored on KU Leuven maintained drives which are automatically backed up.

Maaike Vandermosten is responsible for availability of the data storage during and after the project.

The scripts to generate the results are saved through GitLab and maintained in a specific repository managed by a subset of researchers involved in the project. A code respository will be created for each data type (behavioural processing, EEG processing and MRI processing).

Who will manage data preservation and sharing?

Maaike Vandermosten is responsible for ensuring the availability of data storage during and after the project, as well as for overseeing the sharing of the data if it is made publicly available.

Who will update and implement this DMP?

Marlies Gillis will be responsible for updating and implementing the DMP.