DMP title

Project Name IDN/21/010 (C1-C2-IDN DMP) - DMP title
Project Identifier IDN/21/010
Grant Title IDN/21/010
Principal Investigator / Researcher Joost Vennekens
Institution KU Leuven

1. General Information

Name of the project lead (PI)

Joost Vennekens

C1-C2 Project number & title

IDN/21/010 - Computationele Modellering van Sociale Cognitie en geassocieerde Deficits door middel van Artificie?le Neurale Netwerken.

2. Data description

2.1. Will you generate/collect new data and/or make use of existing data?

- Generate new data

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

WP1: Data set of images scraped online

Type of data	Format	Volume	How created
Images	.jpg, .jpeg, .png	100's GBs	Scraped by web searches
Annotations	.txt	KBs	Annotated by researchers
Program code defining the image scraper	.py	KBs	Programmed by researcher
Program code defining the web interface used for annotating the scraped images	.py, .html, .js	KBs	Programmed by researcher

WP2: Code

Type of data	Format	Volume	How created
Program code generating ANN, performing data analysis...	.py	KBs	Programmed by researcher
Trained ANN models	binary	GBs	Trained by Machine Learning

WP3: Validation using fMRI

Type of data	Format	Volume	How created
brain imaging data	.dcm	TB	generated by MR-scanner
neuropsychological data	.txt	KBs	generated by psychophysical software

WP 4+5: the same as WP2

3. Ethical and legal issues

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

Personal data relating to study participants including name and date-of-birth will be collected for ID purposes during data collection. This information will only be available to researchers directly involved in recruitment, screening, and planning of data collection (e.g. MR scanning and neuropsychological assessments). For the remainder of the study, all derivative data will be coded and thus pseudonymized. The file linking the code and personal identifiers age/dob will only be accessible to authorized individuals and stored in restricted access, secure environment managed by the KU Leuven/UZ Leuven ICT facility.

3.2. 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, add the reference to the formal approval by the relevant ethical review committee(s).

This project involves human participants including healthy volunteers and hospital in- and out-patients with frontotemporal dementia or autism spectrum disorder. It will be submitted to the UPC-KU Leuven local ethics committee and the UZ Leuven EC.

3.3. Does your research possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

Yes, in the long term. Legal aspects regarding third-party involvement are covered in agreements developed by KU Leuven, LRD.

3.4. Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions regarding reuse and sharing are in place?

No.

4. Documentation and metadata

4.1. What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?

-) Image dataset: a general description of the images will be provided, such as the folder structure in which they are stored, the dimensions and image types considered, if there are duplicates present, etc.

-) Trained ANN models: a description of which weights file corresponds to which type of model (i.e., which architecture), as well as the values of the hyperparameters the model was trained with, will be provided.

-) Python code: the code will be documented inline. This inline documentation can be transformed into external documentation by means of tools like Sphinx (https://www.sphinx-doc.org/en/master/). The code will contain a Sphinx setup, so that this type of external documentation (typically in HTML format) can be generated from scratch if desired, as well as include a copy of such external documentation.

The codebook will contain information on study design, sampling methodology, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively. 2. Research methods and practices (including the informed consent process) will be fully documented. Details on the setting of the data collection, the selection of participants and the instructions given to researchers will be documented. Any auxillary data relating to data collection e.g. example neuropsychological assessment forms, will be added to the documentation, as well as an overview of all steps taken to remove direct identifiers in the data (e.g., name, address, etc.).

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

Metadata of the numerical dataset will be created manually based upon commonly used terminology in the fields of neuroimaging, psychiatry and biostatistics. All metadata pertaining to ANN research will be provided in the form of simple text (*.txt) or markdown (*.md) files.

For brain imaging data we will work with the BIDS format, which standardizes directory structure and additional metadata.

5. Data storage and backup during the C1-C2 project

5.1. Where will the data be stored?

The data pertaining to one of the PhD candidates -- Laurent Mertens -- who is responsible for the image scraping, image annotation interface and development and training of ANNs, is stored on three separate systems:

1) a work laptop on which the code is developed,

2) a workstation located in his office in De Nayer, on which the code is deployed and all scraped images are also backed up,

3) a server located in a private space, on which the annotation interface is deployed and which also contains a copy of all scraped images.

Master copies of the neuroimaging and neuropsychological data will be kept on our secure research unit central storage facility, and time-version stamped. Copies can be made and kept on personal devices in accordance with the level of authorisation of the user and the data security level of their device. 2. Sensitive personal data concerning the study participants will be stored in a KUL/UZ secure environment. 3. MR imaging data will be stored on the UZ Leuven data drive, and reconstructed files will be transferred to the hospital PACS system. We will use KUL/UZ managed storage and file-sharing facilities as well as the REDCap platform for active use of the data during the project.

5.2. How will the data be backed up?

For Laurent Mertens, systems 1 and 2 (see 5.1.) have a daily backup routine on a separate HD in place. For system 1, this is an external HD, for system 2 this is an additional internal HD.

All code is also pushed to a GitLab repository, and so can be retrieved, even in case all 3 systems and their backups should fail. Furthermore, a manual backup on an external HD of all scraped images and other data (e.g., trained models) is also taken from time to time.

The neuroimaging and neuropsychological data will be stored on the university's and UZ Leuven central servers with automatic daily back-up procedures.

5.3. 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.

A request for additional data storage and back up will be requested from the university ICT services.

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

1. The expected costs for data storage and back up (REDcap, KUL, UZ data) are estimated to be up to €2000 euro per year for 5-10TB. 2. Part of the allocated project budget will be used to cover the costs for storage and backup.

5.5. Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

For Laurent Mertens, the backups are encrypted and password protected. To access the original data (i.e., not the backed up copies), one would need access to the systems. This can be either digitally, which would require knowledge of the IP adresses and login credentials, or physically. It is the candidate's responsibility to make sure these credentials are not only kept safe but of sufficient strength to deter any unwanted data manipulation.

The identifiable neuroimaging and neuropsychological data files from this study will be managed, processed, and stored in a secure environment (KUL/UZ)

6. Data preservation after the end of the C1-C2 project

6.1. Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is

the case (legal or contractual restrictions, physical preservation issues, ...).
We expect all data to be kept.

6.2. Where will these data be archived (= stored for the long term)?
The data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?
The data that will be compiled to realise the project objectives will be hosted on the servers of KU Leuven. In view of the expected size of the dataset (5 TB), estimated cost will be 783 euro pa * 10 years = €7830

7. Data sharing and re-use
7.1. Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions or because of IP potential)?
The project will make use:

1) of a large dataset of images representing various people scraped from the internet. Although these images are all publicly available, they cannot legally be redistributed (GDPR, copyright...). What can be readily shared though, is the list of URLs corresponding to all used images.

2) of a set of fMRI scans of human brains. Pseudonimyzed data can be made available for further analysis in line with the terms of the ICFs and following advice from the relevant local ethics committees and LRD.

7.2. Which data will be made available after the end of the project?
The type of data that will be made available will be determined on an ad hoc basis and in adherence with the informed consent of the participants and advice from the relevant ethics committees and LRD.

7.3. Where/how will the data be made available for reuse?

- In a restricted access repository
- Upon request by mail
- Other (specify):

Different data types will be made available by different means according to the nature of the data and request. For example, pseudonimyised pre-processed brain scans may be shared with collaborators for new analyses or uploaded to a discipline specific repository (eg Neurovault), and speadsheets with pseudonimyised data can be made available on request. Publications relating to the study will be provided in open access via Lirias 2.0.

7.4. When will the data be made available?

- Upon publication of the research results

The data derived from this study has multiple end-points and is part of a longitudinal research programme that will include related data at a future time-point. When it will be made available is therefore dependent on the type of data and how it will be analysed. The initial results of the first cross-sectional analyses will be made available in abstract form at the earliest opportunity.

7.5. Who will be able to access the data and under what conditions?
Access will be granted upon written request to the creators of the dataset. Commercial reuse is not allowed.

7.6. What are the expected costs for data sharing? How will these costs be covered?
The major cost of data-sharing will be long-term large volume storage after completion of the project. These costs will be covered by part of the allocated budget.

8. Responsibilities

8.1. Who will be responsible for the data documentation & metadata?

Each PhD candidate will be responsible for his/her own data.

8.2. Who will be responsible for data storage & back up during the project?

Each PhD candidate will be responsible for his/her own data.

8.3. Who will be responsible for ensuring data preservation and sharing?

This responsibility will lie with the supervisor, as this requires a long term commitment most likely extending beyond the tenure of the PhD candidates.

8.4. Who bears the end responsibility for updating & implementing this DMP?

The end responsibility for updating and implementing the DMP is with the supervisor (promotor).

To this end, the PhD candidates commit themselves to keeping the supervisor informed about the status of their respective data backup and preservation.