LEIBNIZ-INSTITUT FÜR NEUROBIOLOGIE MAGDEBURG

cni combinatorial neuroimaging

CNI Research Data Management for Human Imaging Studies

Study Management

A study is defined as coherent experiments performed on a number of subjects while applying identical methods (to be described in the methods section of publications). A study may include pilot experiments where methods are optimized until the final experimental setup and design is reached.

To conduct experiments at CNI imaging facilities, each study must be registered via the website www.lin-magdeburg.de/cni to provide initial information about the responsible principal investigator, scientist in charge of performing the study, required imaging modalities and planned number of measurements.

CNI staff members will check the technical feasibility of the study and support the scientists in defining and documenting the experimental setup and design to start the study.

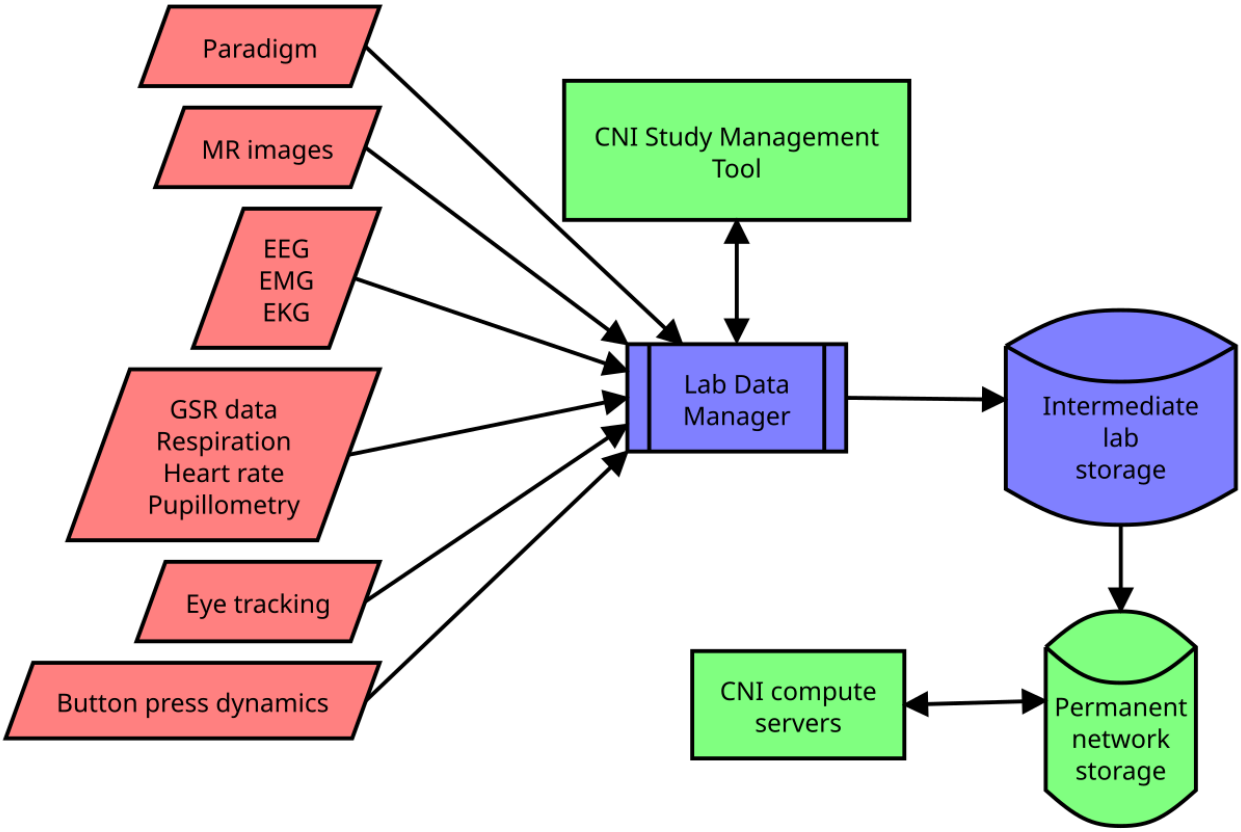
The required measurement time will be organized via the open source eGroupware web application, that has been adapted to serve as CNI's study management tool (SMT) in which each measurement time has to be linked to the respective study. In addition, the booking title can be used to specify the subject-ID and more specific description of the status of the measurement (e.g. pilot measurement, control/test group). Any change of a calendar entry is documented. This procedure ensures an overview of all planned and conducted measurement sessions of a study.

Experimental setup and data acquisition

CNI's human imaging labs comprises complex setups of multiple custom and commercial devices for stimulus presentation, feedback and intervention as well as neural, psychophysiological and behavioral data acquisition. This includes anatomical and functional MRI, electroencephalography, electrocardiography, facial electromyography, skin conductance, heart rate, respiration, pupil dilation, eye tracking, response times & button press dynamics, and video recordings.

Since each device uses its own file format and filename conventions, manual storage of experimental data is prone to errors. Therefore, storage of experimental data is performed in a semi-automatic and assisted way including sanity checks based on study-specific details (i.e. names, number, format and size of files) and supplemented by session-specific remarks added by the user manually or via an electronic lab-book. For permanent storage and data analysis, all acquired data is compiled in a well-organized directory structure.

The data collection network architecture is depicted in the schema below. The central component is the Lab Data Manager (LDM) that collects the data from the different data acquisition devices via a network switch. The current hardware and installed software of the multiple computer devices that handle stimulus presentation and data acquisition is documented in a lab-configuration status file stored and kept up-to date at the LDM. CNI's study management tool (SMT) provides additional information about session date and subject IDs. The LDM finally checks, organizes and stores all source data of a session (experimental data and session specific notes) on the intermediate RAID-storage located within the lab and to the permanent network storage (PNS) located outside the lab. Here the data is accessible in a read only mode for authorized persons. CNI provides compute-servers with RAID systems to perform fast and secure data analysis.



Network architecture for data acquisition, management, and storage.

Source data organization and storage

For usability and archival, the acquired data is stored in a well-defined structure adapted from the community-developed BIDS (Brain Imaging Data Structure) standard. Essentially, it provides a common directory structure with a specific naming scheme for files and sub-directories. We implement this with one directory per study with sub-directories for each measurement session, which contain all acquired data. Furthermore, we augment this structure with a few important meta information:

1. Study description containing the selected data modalities, technical specifications of data acquisition, hardware and software.
2. Checksums (e.g., SHA256) for all files in a session sub-directory ensuring data integrity.
3. Session notes of potential relevance for data quality and data analysis.

The naming schemes follows the definition below. Brackets indicate placeholders, *n* is the subject number and *yyyymmdd[HHMM]* the time stamp where hour and minute is optional.

```
- {study-code}/
  - {study-code}_description.xlsx
  - {subj-id-1}_{yyyymmdd[HHMM]}/
    - {subj-id-1}_{yyyymmdd[HHMM]}_Paradigm.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_MRI.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_EEG.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_ResponseEvents.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_ResponseDynamic.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_Physio.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_Video.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_EyeLink.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_Questionaire.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_Interview.tar.gz
    - {subj-id-1}_{yyyymmdd[HHMM]}_Notes.txt
    - {subj-id-1}_{yyyymmdd[HHMM]}_Checksum.txt
  - ...
  - {subj-id-n}-{yyyymmdd[HHMM]}/
```

Source data for human imaging are stored at linstore CNI/HBI/Source_Data/{study-code}

This data structure implements multiple levels of data safety. First, data integrity is ensured by a checksum for each file to test against. In turn, as there is a checksum for each file, also the completeness of data acquired in a measurement session can be verified. Moreover, the file names lack spaces and special characters for strong compatibility across file systems and operating systems. Finally, file names are specific and can be assigned to its corresponding measurement session. Ultimately, all data is saved on an intermediate lab storage device as well as on a network storage with regular backups and archival, so that no data is lost.

The data will be provided to the PI of the study for analysis and publication. CNI offers compute servers for data analysis. Data analysis pipelines, analysis software, derived data, final results and meta-data that is necessary to reproduce the results described in the publication should be stored together with the source data of the study as soon as possible. This way, all data can be made accessible in open data repositories.