ZAT: Guidelines for Research Data in Research and Development Processes of Pro-adaptive Cognitive Assistive Technologies (Pro-CAT)

Anne Ferger, Robin Grashof, André Frank Krause, Thomas Schmidt, Jordan Schneider, Sinan Yavuz
(alphabetical order)
Whitepaper

10.5281/zenodo.15187761

1. Context

This section gives context about the project *Center for Assistive Technologies Rhein-Ruhr* (Zentrum Assistive Technologien Rhein-Ruhr, ZAT[1], see Wild-Wall et al. 2024) in which this whitepaper was written, and in introduces concept of *pro-adaptive cognitive assistive technologies* (pro-CAT) which is central to the ZAT project.

Systems including pro-CAT offer cognitive assistance which dynamically adapts to the participant's needs by using artificial intelligence to predict individual support requirements. It automatically adapts the type of assistance and communication offered to the individual participant and the respective application context, taking into account predictive changes, e. g. expected course of illness, imminent changes in the environment or expected learning progress. If deterioration is expected, support can thus be increased, but expected improvements can also be reduced in order to maintain the participant's independence for as long and as extensively as possible.

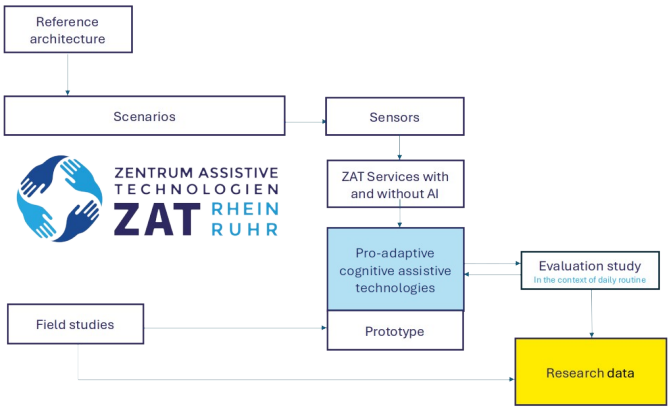

Figure 1: The role of research data in ZAT project

Our project ZAT pursues the goal to develop and research on pro-CAT. Here we build a reference architecture for assistance systems including pro-CAT. We will collect data

1 https://zat.nrw/

in different scenarios using various sensors to train the artificial intelligence for our systems, develop and apply base services, test the systems in real settings and publish open data for research purposes. Figure 1 illustrates where in the ZAT project research data is collected and processed. In specific application scenarios, many different time series data must be used to enable analyses for different sources, which lead to an individual adaptation of the assistance of the pro-CAT. For example, acceleration data from wearables, EEG data, position data and audio and video data with time stamps that are to be synchronised with each other are conceivable here. This allows different data from multiple sensors to be recorded and analysed simultaneously or in response to a specific event. Therefore, time series data play a particularly important role in pro-CAT.

In a complex context like developing and doing research on cognitive assistive technologies, interdisciplinary teams are required. Therefore a common terminology is required to facilitate interdisciplinary research. Explicit guidelines can ensure traceability and transparency, both of which are important factors for the publication and promotion of open data for research purposes. This can include the availability of sensor information as metadata or the corresponding schema files with which the data are validated.

The following guidelines aim to define a basic terminology to facilitate the interdisciplinary exchange and a common understanding in ZAT, without trying to prescribe the chosen alternatives of specific terms for the respective disciplines. We also provide a classification of relevant data types and propose measures for standardization and validation of data formats relevant for the research in ZAT.

2. Terminology

This section introduces some basic terminology, meant to serve as a common ground for the following sections and as a starting point for categorising research data created in individual scenarios. The terminology does not claim to be exhaustive, it is not meant to be prescriptive, and we are aware that some of the chosen terms may be used in significantly different meanings in other disciplines. The scope of the terminology is thus restricted to the questions of data organisation treated in this document.

The first group of terms is concerned with the basic abstract entities for speaking about research data:

We call an **event** (alternative terms: *session*, *communication*, *interaction*, *data set*, *trial* / de: *Ereignis*, *Datensatz*) the actual, individual occasion which is observed/recorded for research purposes. Examples of events are: an interactive session with an assistive technology, an individual instance of an experiment, or an interview conducted with participants of an experiment. Individual research

data files will, as a rule, pertain to one and only one event.

The term event (and event marker etc.) is also used in neurophysiological data in a different meaning. In speech transcription (EXMARaLDA[2]), it refers to an individual time-aligned annotation.

We call a **participant** (alternative terms: *speaker*, *informant*, *test subject*, *agent*, *user*, *patient* / de: *TeilnehmerIn* / *SprecherIn*, *InformantIn*, *Testperson*, *ProbandIn*, *NutzerIn*, *PatientIn*) any person taking part in an *event*. Besides test subjects proper, this also encompasses researchers actively participating in events. An assistive technology (e.g. a robot) can also be modeled as a participant.

We call a **collection** (alternative term: *study*, *data set* / de: *Kollektion*, *Datensatz*) a systematic, completed compilation of research data files, organised according to the *events* they belong to.

We call a (**linguistic) corpus** (de: das *Korpus*, pl. *Corpora/Korpora*) a *collection* in which language data has been processed (i.e. consistently transcribed and annotated, represented in structured, machine-readable files) in such a way that it can be used in a corpus linguistic research approach.

The concepts of collections, events and participants (or the alternative terms) are often used as a high-level structural distinction in corpus management tools and archives. The distinction between collection and corpus is also reflected in the distinction of different "data maturity levels" as described in Wamprechtshammer et al. (2022).

The second group of terms serves to distinguish research data according to the stage and the way in which they are created:

We call **raw data** (alternative term: *original data* / de: *Rohdaten* / *Originaldaten*) the observational data as it is recorded in the *event* itself. The form of a piece of raw data is determined by the device with which it is recorded, e.g. a video camera, a heart rate sensor, a motion capture suit, an eye tracking device, etc. Because of this, raw data often comes in proprietary and/or binary formats. Raw data may be used within the project that created it (e.g. if eye tracking data is further processed in the software that comes with the eye tracker) but it is usually not suitable for archiving and reuse.

Primary data (alternative term: *working data* / de: *Primärdaten* / *Arbeitsdaten*) is *raw data* transformed, without (relevant) information loss, to a non-proprietary, openly documented format, ideally a standardized format. For many types of data, primary data will be in text form (esp.: CSV and XML, see section 4). For audio, video and image data, the relevant standards are binary formats (esp. WAV, MPEG-4, JPG, see section 4). All primary data should be described with adequate *metadata* (see below) and preserved/archived for reuse. Frequently, the process of transforming raw into primary

2 https://www.exmaralda.org

data includes the synchronization of different pieces of *time-series data* (see below).

We call **derived primary data** (de: *abgeleitete Primärdaten*) data which constitute an excerpt or a rearrangement of *primary data* without adding any information to it. Examples for derived primary data include: the soundtrack of a video file extracted as a WAV file, face recognition data extracted from a video file, images from different video files arranged into a 4-in-1-setup in a single file, or F0/F1 frequency data extracted from an audio file. As *derived primary data* add no or little information to the *primary data* they are based on, they can theoretically be reconstructed at any time from the *primary data* without major costs. They may therefore be excluded from preservation/archiving if, for instance, storage space restrictions mandate this.

Secondary data (de: *Sekundärdaten*) is analytical data referring to *primary*, *derived primary*, or other *secondary data*. They are added manually, i.e. in an intellectual process guided by research methodology, automatically with some non-trivial technique, or in an interplay between automatic and manual methods (e.g. when the output of an automatic speech recognizer is corrected and refined by a human transcriber). Processes for creating secondary data are very diverse and run under different names, most importantly:

- **Transcription** (de: *Transkription*), meaning the transfer of spoken language to a systematic written representation, often including time-alignment of parts of the transcribed text, and their assignment to participants.

- **Labelling** (alternative term: *coding* / de: *labeln/kodieren*), meaning annotation with a predefined (closed vocabulary) set of text labels, for example, lemmatisation or part-of-speech-tagging of a transcript, or the labeling of objects/persons that gaze is directed to (as recorded by an eye-tracker).

- **Annotation** (de: *Annotation*), meaning any kind of analytical/categorial information added to the primary data (subsuming transcription and labeling), for example, the free description of body movements visible in a video.

Metadata (de: *Metadaten*) serve to systematically describe *collections*, *events*, *participants* and individual pieces of *primary* and *secondary data*. Examples of metadata include:

- Name and general description of a collection, names of the researchers responsible for it (cf., for example the Dublin Core metadata set: https://www.dublincore.org/)

- Time and place where an event took place, specifics of the equipment used to record it

- Age, gender, other sociobiographical data about participants

- Camera perspective, resolution of a video file
- Transcriber, transcription convention used for a transcript
- Annotator, annotation guideline used for labeling a piece of sensor data

As explained in detail in Schmidt (2022), metadata serve different functions in the research data life cycle: they are crucial as information for the research process itself, but also for making archived data findable and reusable.

Further terms in need of a definition for the purpose of these guidelines are:

Data model (de: *Datenmodell*), meaning an abstract description of data objects and the relations between them.

Data format (de: *Datenformat*), meaning the specification of the concrete digital representation of a data object (its "serialization") such as XML, CSV, MP3.

Schema (de: *Schema*), meaning a formal description of a data format.

Validation (de: *Validierung*), meaning a procedure to determine whether or not a given piece of data (a file) adheres to the specification of its data format.

Time series data (de: *Zeitreihendaten*), meaning any data consisting of sequential observations where each observation is associated with a timestamp. Audio recordings, video recordings, most sensor data, and any time-aligned transcript are all examples of time-series data

Synchronization (de: *Synchronisierung*), meaning the process of relating different pieces of *time series data* of the same *event* to one common timeline (so that any timestamp x in any file refers to the same absolute point in time). Synchronization can either be done in the recording situation itself, or it can be brought about retroactively when raw data are transformed into primary data.

Temporal resolution (alternative term: *sampling rate* / de: *zeitliche Auflösung* / *Samplingrate*), meaning the granularity with which time-series data are recorded. For example:

- Video is typically recorded with 25 or 50 frames (=single images) per second (fps)
- Audio is typically recorded with 44.1kHz or 48kHz, i.e. with 44100 or 48000 samples per second

- Sensor data can have very different resolutions. For example, eye tracking data can be recorded with as temporal resolution up to 1200Hz, but usually much lower rates are used.

3. Classification of Research Data Types

To organise and document a collection or a corpus and prepare it for reuse, it is crucial that all objects contained therein are described with suitable metadata, and that relations between different objects are clearly defined. We propose the following basic data types, as illustrated in Figure 2:

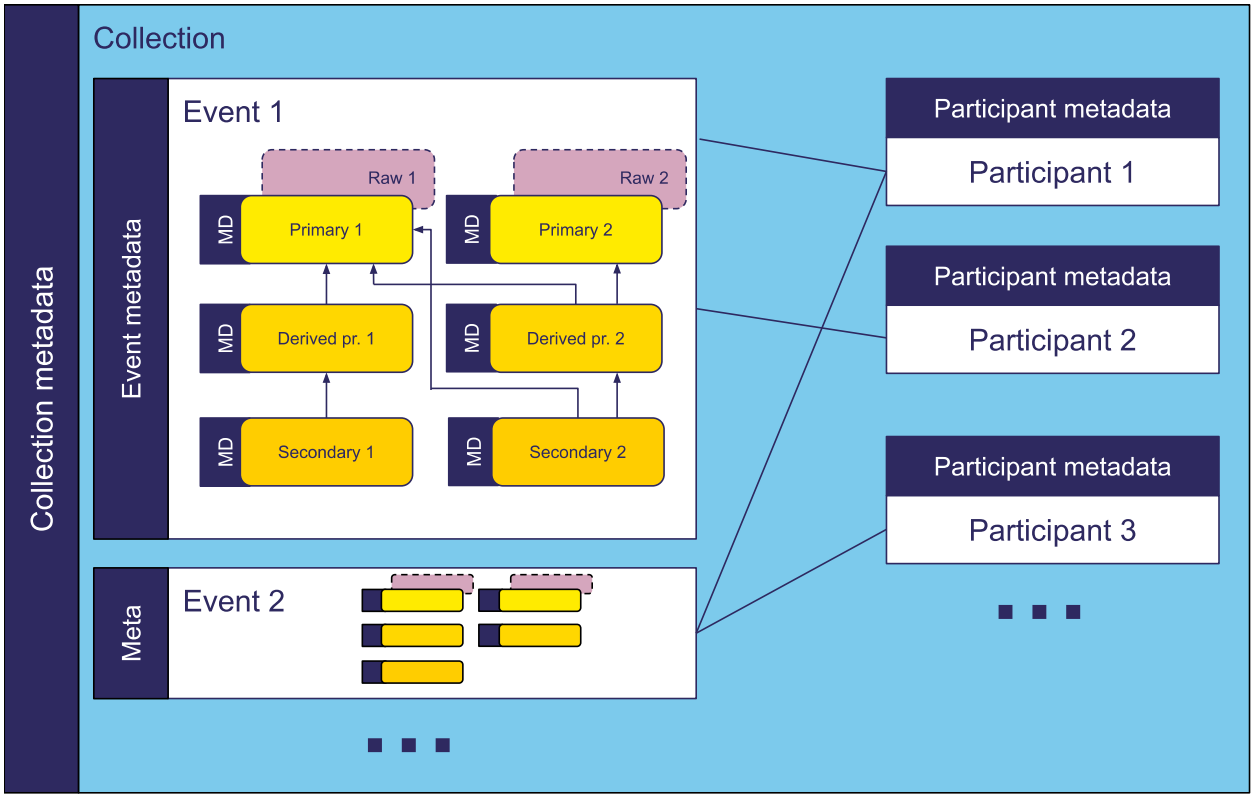

Figure 2: Overview of data types - "MD" is short for "metadata", "Derived pr." for "Derived primary data", "Primary" for "Primary data", likewise for "Raw" and "Secondary"

- **Collections** (Corpora), **Events** and **Participants** as described above. A collection is made up of one or more events. Events and participants. Events and participants are assigned to each other in a many-to-many (n:m) relationship. Properties of collections, events and participants are described with suitable metadata, where some properties (such as: time, place, gender) will be relevant and meaningful for all collections, while others (such as: role) may be collection-specific.

- **Primary video data**: Each video file will typically be assigned to exactly one event. It may also be assigned to one or more specific participants, and to the corresponding raw data object. Properties of the video include technical parameters (such as: recording device, frame rate or image resolution) as well as information on its content (such as: which perspective was recorded, which part of an event is captured). any

  The following is an example for technical parameters of a typical MPEG-4 video file:

Attribute	Value
Format/Container	mp4/AVC
Encoding	H.264/MPEG-4 Part 10 AVC
Resolution	1980x1080
Frame rate	25 FPS
Bit rate	up to 5000Kbps
Bit rate type	constant
Scan type	progressive
Audio format	AAC
Audio sampling rate	48kHz
Bit rate	384Kbps
Audio channels	Stereo
Recording device	Sony Alpha 6700

  Relations of this video file to other objects could be described as follows:

Attribute	Value
Filename	Event_001_Video_003.mp4
Raw file	Event_001_Video_003.MTS
Event	Event_001

  Further, content-related properties for this video file could look as follows:

Attribute	Value
Perspective	Close-up of robot
Position	Static, on a tripod in front of the table

- **Primary audio data**: Each audio file will typically be assigned to exactly one event. It may also be assigned to a specific participant (for instance, when it is recorded by a lapel microphone), and to the corresponding raw data object. Properties of the audio include technical parameters (such as: recording device, sampling rate, bit rate, number of channels) as well as information on its content (such as: which perspective was recorded, which part of an event is captured).

- **Primary image data**: Each image file will typically be assigned to exactly one event. It may also be assigned to one or more specific participants. Properties of the image include technical parameters (such as: recording device, frame rate or image resolution) as well as information on its content (such as: which perspective was recorded, which part of an event is captured).

- **Primary sensor data**: Sensor data as it is produced by one sensor (such as: heart-rate, temperature) will typically be assigned to exactly one event. It may also be assigned to one or more specific participants. Properties of the sensor data include technical parameters (such as: sensor device, temporal resolution) as well as information on its content (such as: which participant was recorded, which part of an event is captured).

- **Primary logging data:** Logging data as it is produced by one assistive system will typically be assigned to exactly one event. It may also be assigned to one or more specific participants. Properties of the sensor data include logging parameters (such as: logging device) as well as information on its content (such as: which system was recorded, which part of an event is captured).

- Primary **questionnaire data:** Questionnaires are lists of questions asked to participants that go beyond simple participant metadata. They are modeled as objects in their own right and typically assigned to exactly one participant (but not necessarily to an event).

- **Derived primary data:** Each object of derived primary data is related to one or more primary data objects. This relationship, as well as the procedure of how the data was derived, must be documented. This could look as follows for a 2-in-1 video:

Attribute	Value
Filename	Event_001_DerivedVideo_001.mp4
Derived from	Event_001_Video_003, Event_001_Video_004
Type	2-in-1-Side-By-Side

Technical metadata for derived primary data should be documented in the same way as for

primary data (see above).

- **Secondary data:** Each object of secondary data is related to one or more (derived) primary data objects. This relationship must be documented as well as the guidelines (e.g. transcription conventions), vocabularies (e.g. tag sets) and tool(s) (e.g. a text editor, a specialised annotation tool) used. In practice, it is common to have several types of secondary (e.g. a transcription of verbal behaviour, a description of body posture, and part-of-speech-tags) in a single file (e.g. an EAF file from the ELAN tool).

Attribute	Value
Filename	Event_001_Transcript_007.eaf
Primary data	Event_001_Video_003, Event_001_DerivedAudio_003
Transcription convention	GAT2
POS tagset	STTS 2.0
Transcription tool	EUDICO Linguistic Annotator (ELAN), v6.2
POS tagger	TreeTagger, v3.2.5
Transcriber	Franz Gans

---

4. Measures for Standardization, Consistency and Validation of Data Formats

Collected relevant research data types in ZAT include sensor data such as heart rate, temperature, EEG, video and audio data as well as questionnaires. For these examples, which will be part of collections or corpora in ZAT, we propose measures for research data management such as standardizing, ensuring consistency and validating.

Research Data that is relevant within the context of a research question will be processed through various stages after it has been collected and stored. This typically includes planning data collections, producing data collections including metadata, processing the data to

Figure 3: Research Data Life Cycle (see also RDMkit)

ensure data quality and correct any issues, analysing and interpreting the data using varying methods, preserve the data and finally reuse it (see e.g. *RDMkit*[3]). An example of typical research data life cycle is shown in Figure 3.

Specifying and documenting data formats in the context of open research data is crucial to ensure collections or corpora can be maximally useful for other researchers, who might wish to integrate the collected data into their own research workflows. The collections or corpora should fulfill the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016), so they are **f**indable, **a**ccessible, **i**nteroperable and **r**eusable. *Findability* can be enhanced by using standardized metadata which is made available using catalogues[4], so it can be found by searching for its metadata. *Accessibility* is heavily depending on the repositories (storage locations or archives for digital objects that make them available to a public or restricted group of users) or archives where the resources are shared. *Interoperability* and *reusability* can be improved by using standardized formats and documenting how the respective files are following these standards.

Additionally, it is important to ensure that datasets are of a high consistency to avoid the problem of Garbage in, Garbage out (GIGO). That is, for the analysis to yield accurate and correct results, the collected data must be as consistent as possible. This is particularly relevant in the field of research for pro-CAT systems, where the datasets are used to train models and algorithms that make decisions that could have a significant impacts on participants. Possible inconsistencies in the data could be using different export formats for the same sensors, so they are not easily comparable, spelling errors in filenames or metadata, so a search for a specific term does not yield complete results, or using different number formats for the same types of data values. Additionally to data consistency, data quality is important for accurate and high-quality research results. The discussion of data quality (see e.g. Hedeland (2020) on data quality of audiovisual language resources) goes beyond the scope of this whitepaper, thus in the following we focus on data consistency. To achieve higher and check data consistency, the use of standardized formats for open research data is essential. It is still common for research data files to be published in inconsistent, poorly documented and unspecified ad-hoc data formats, such as a collection of undocumented CSV files.

To automatically check data consistency, validation can be performed against the definition or standard of what is considered consistent data. Following examples of inconsistencies can be checked using validation:

- **Data format:** Checks whether the data conforms to the expected format
- **Value number format:** Checks whether the values in the data conform to the expected format to the specified data types, e. g. floating point, character etc.
- **Spelling of values:** Check if spelling of identical values diverges.

3 RDMkit: The ELIXIR Research Data Management toolkit for Life Sciences URL: https://rdmkit.elixir-europe.org

4 Such as the Virtual Language Observatory: https://vlo.clarin.eu

- **Missing values:** Checks whether the data conforms to the expected format, e. g. separation of data by commas and to the specified data types, e. g. floating point, character etc.
- **Range limit:** Checks whether the data is within a specified range.
- **Temporal resolution:** Checks wether data was recorded with the established temporal resolution
- **File naming:** Checks if file names adhere to naming schemes for files (for example no whitespaces in filenames).

Defining what is considered consistent data can be achieved using schemas. These schemas should be derived based on evaluation criteria after analysing the requirements of the study outcomes (Mahesh et al. 2019). These definitions, standards or schemas can also be used as a way of documenting the data, as they define and explain the structure and content of a data set. Below, we propose standardised schemas for selected data types which are relevant for the development of pro-CAT systems, namely exports from various wearables, EEG data, metadata, transcriptions, annotations and audio and video data. For these different types and formats of research data, different kinds of schemas are required, as well as different validation mechanisms. Therefore we list the different kinds of schemas with their respective validation mechanisms we propose below.

Wearable sensors

For the sake of simplicity we work with CSV outputs directly from wearable sensors (such as heart-rate sensors). For these CSV-based files we propose to use the open source CSV Schema[5] created by the National Archives[6]. There are tools to validate these schemas published with the schema documentation as well[7]. For ease of use and automated data validation in the ZAT project without having to rely on technical knowledge, we developed a CSV validation software that can be used to create CSV schemas and validate data[8].

EEG files

While we still need to gain more experience working with EEG files from a research data perspective, we currently propose following the EEG-BIDS, an extension to the brain imaging data structure for electroencephalography (Pernet et al. 2019)[9]. We will work on proposing validation mechanisms for EEG files in the future.

Metadata, transcriptions and annotations

For XML-based files (such as metadata files and transcriptions files (EAF, TEI)), XML schemas can be used for validation. There are official schemas for EAF files and TEI files following the ISO standard for

5 https://digital-preservation.github.io/csv-schema/
6 https://www.nationalarchives.gov.uk/
7 See https://github.com/digital-preservation/csv-validator
8 This code can currently be shared on requests.
9 https://bids-specification.readthedocs.io/en/stable/modality-specific-files/electroencephalography.html

the transcription of spoken language („ISO 24624:2016 Language resource management — Transcription of spoken language“, based on Schmidt (2011)). We propose to use existing XML schema validation tools continuously on research data using version control and continuous integration functionality from e.g. GitLab.

Video and Audio files

For audio and video files we propose specifying the technical parameters similar to the example in the research data type *Primary video data* above. Video files can exhibit a range of errors due to bugs in the encoding software, file corruption during transmission or non-compliance to video encoding standards. Yet, to the best of our knowledge, there are no open source tools that test mpeg-4 video files for strict compliance to the ISO-standards.

Common problems can be captured using ffmpeg by issuing ffmpeg to decode the full video and report errors:

```
ffmpeg -v error -i testfile.mp4 -f null -
```

Issues with the mpeg-4 meta-data of a video-file can be checked using the tool ffprobe (part of the ffmpeg suite). ffprobe extracts information from a wide range of audio – and video file formats, e.g. used codecs, bitrates, durations, resolutions, sample rates, and other technical properties.

```
ffmpeg -i testfile.mp4
```

A specialized commercial tool to check the validity of a mpeg-4 video is provided by Jongbel Media Solutions[10]. The MPEG-4 Video Validation Module included in their software product provides "validation of MPEG-4 video elementary streams according to the ISO/IEC 14496-2 standard."[11]

Audio files can be validated in a similar way using for example flac:

```
flac -t
```

5. Outlook

As these guidelines are used in practice by working with them in the ZAT project, we plan to update this document and publish newer versions with our experiences. We will also reevaluate the data types relevant to ZAT and add further schema and validation proposals.

10 https://www.jongbel.com

11 https://www.jongbel.com/automated-validation/mpeg-4-video-validation/

References

Hedeland, H. (2020). Towards comprehensive definitions of data quality for audiovisual annotated language resources. In CLARIN Annual Conference Proceedings, 2020 (pp. 93-103). ISSN 2773-2177 (online). Eds. C. Navarretta and M. Eskevich. Virtual Edition, 2020.

Mahesh, B., T. Hassan, E. Prassler and J. Garbas. (2019). Requirements for a reference dataset for multimodal human stress detection. *Emotion Aware Workshop* held in conjunction with PerCom 2019. https://ieeexplore.ieee.org/document/8730884

Pernet, C. R., Appelhoff, S., Gorgolewski, K.J., Flandin, G., Phillips, C., Delorme, A., Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. Scientific data, 6 (103). doi:10.1038/s41597-019-0104-8

Schmidt, T. (2022). Daten und Metadaten. In M. Beißwenger,L. Lemnitzer, Lothar & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 249-258). Wilhelm Fink.

Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. Journal of the Text Encoding Initiative 1 (2011). doi:10.4000/jtei.142.

Wamprechtshammer, A., Aznar, J., Arestau, E., Hedeland, H., Isard, A. et al. (2022): QUEST: Guidelines and Specifications for the Assessment of Audiovisual, Annotated Language Data. , 8, pp.90, 2022, Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology, Kristin Bührig, 978-963-306-910-3.

Wild-Wall, N., Ressel, C., Kannen, K., Krause, A. F., Büscher, S., Mosler, B. und Arntz, B. (2024). Strukturen zur Berücksichtigung ethischer Aspekte in der Entwicklung von digitalen assistiven Technologien für vulnerable Gruppen. https://www.nomos-shop.de/de/p/wer-rettet-die-welt-gr-978-3-7560-2352-3

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18 .