DATA MANAGEMENT PLAN (DMP)

For COST Action “MultiplEYE” (CA21131)

Date of creating this version (dd/mm/yy):

20/03/23 (version 0.1)

Date of last update:

09/01/24 (version 1.0)

Authors:

Marie-Luise Müller, Leibniz Institute for Psychology

Deborah Jakobi, University of Zurich

Ramunė Kasperė, Kaunas University of Technology

Nora Hollenstein, Kaunas University of Technology & University of Zurich

Lena Jäger, University of Zurich

Contact:

Lena Jäger, jaeger@cl.uzh.ch

Project Details

The MultiplEYE COST Action (*MultiplEYE*: Multilingual Eye-Tracking Data Collection for Human and Machine Language Processing Research) aims to foster an interdisciplinary network of research groups working on collecting eye-tracking data from reading in many languages. The goal is to support the development of a large multilingual eye-tracking corpus and enable researchers to collect data by sharing infrastructure and their knowledge between various fields, including linguistics, psychology, and computer science. This data collection can then be used to study human language processing from a psycholinguistic perspective as well as to improve and evaluate computational language processing from a machine-learning perspective. Furthermore, multilingual eye-tracking data collection refers to the process of

collecting eye movement data from individuals while they read or look at stimuli in different languages. This COST action started in September 2022 and now includes more than 120 researchers and scientists from various areas of expertise (e.g., linguistics, psychology, computer, and information sciences) in 37 countries to support the development of a large multilingual eye-tracking corpus. This will enable researchers to collect data by sharing infrastructure and their knowledge and foster the establishment of common standards for conducting eye-tracking research.

The main outcomes of MultiplEYE will be a large dataset containing eye-tracking data in many languages and a platform for new collaborations leveraging this type of data.

The MultiplEYE data collection will adhere to the FAIR data principles (Wilkinson et al., 2016). FAIR data are data that meet the principles of **findability**, **accessibility**, **interoperability**, and **reusability** (FAIR).

To meet the FAIR requirements, uniform data collection and data processing guidelines defining the management and documentation of the collection procedure and data processing, are described in detail within this data management plan. This includes the following aspects:

- To make the data **findable**, a uniform data naming scheme (including the term “MultiplEYE” as well as a code for language and location of the data collection) will be used for all data, files, and folders created for this action. Moreover, relevant metadata will be provided through a standardized metadata schema which will be shared with every dataset collected for the corpus (for more details, see section 8 of this DMP). When shared through the research data center of the Leibniz Institute for Psychology (RDC at ZPID) and the repository *PsychArchives*, the data will also be identified by a persistent identifier (such as Digital Object Identifier, DOI), so the data can be cited easily and uniquely. Sharing the collected data at the RDC through our own database called the *MultiplEYEStore* will allow us to make the data findable and searchable.
- To make the data openly **accessible**, all data (i.e., eye-tracking data at all stages of the preprocessing pipeline, from raw to aggregated data) will be shared in an open-source repository. This also includes all relevant metadata, code, and software packages for data pre-processing, stimulus preparation and experiment presentation, codebooks,

and additional study documents (e.g., questionnaires). We share all data for public (re)use without any restriction to access (see Section 9 of this DMP).

- Standardized procedures within data collection, (meta)data documentation, and management (including using metadata standards such as the Dublin Core) will allow data exchange and re-use between researchers, institutions, organizations, and countries. Adhering to those standards, which also cover standardization of data processing through a pipeline producing uniform formats, as well as using widely common vocabularies and ontologies, so that data are comprehensible for other disciplines, will increase the data’s **interoperability** (see Section 8).
- One of the main goals of MultiplEYE is to develop a multilingual eye-tracking data corpus, which already presupposes that the data can be reused. To increase the data’s **reusability**, standardized data documentation and management are applied by complying with this data management plan. Moreover, the eye-tracking data will be shared along with all necessary information and metadata enabling sustainable use. Furthermore, clarifying licenses about the data’s usage allowing open and immediate access to the data as well as a transparent data quality control provided by a data quality reviewing team (see Section 7), acknowledging data of any quality level, will enhance the data’s widespread and, therefore, its reuse. The database MultiplEYEStore, hosted by the RDC at ZPID, will provide the appropriate platform for publishing and reusing all collected data.

1. Data Description

MultiplEYE aims to produce a core dataset consisting of eye-tracking data in multiple languages, as well as data collected through a participant’s questionnaire and through comprehension questions. The core dataset will be used for the primary research questions of the project. In addition to this core dataset, collection sites may also choose to collect additional data to investigate certain research questions of interest. These additional datasets will be considered as supplementary to the core dataset and will be subject to the same quality control measures and documentation procedures as the core dataset. It is important to note that the supplementary datasets may not be available at all collection sites, and their inclusion

will depend on the specific research questions being investigated at each site. The final dataset will be made available in a standardized format for (re)use.

The following shows which data is collected and/or created for the core dataset (i.e., eye-tracking data collected through reading experiments plus any other additional data). This applies to all data collection sites. The final dataset (core dataset for eye-tracking data corpus) contains the data in different stages of the process. The table below gives an overview of the different data formats at the different stages:

Table 1: Overview of different stages of eye-tracking data

Name	Data type	Description	Instrument	Format
"[subject_id]_[language]_[session_id]_raw_data_v00.edf"	Eye movement data	Stage 0: Non-human readable data files produced by the eye-tracker. Those files must first be processed by an appropriate tool to make them readable.	Eye tracker	.edf
"[subject_id]_[language]_[session_id]_raw_data_v00.asc"	Eye movement data	Stage 1: Data files converted into a human-readable format. Apart from the encoding, nothing has changed from the first format to the current one.	Eye tracker	.asc
"[subject_id]_[language]_[session_id]_parsed_data_v00.csv"	Eye movement data with x-y-gaze coordinates	Stage 2: Parsed data files that contain one sample (i.e., at least the x- and y-coordinates of one eye and a time stamp) per line. The samples are chronologically sorted.	Eye tracker	.csv
"[subject_id]_[language]_[session_id]_gaze_events_v00.csv"	Eye movement data with gaze events	Stage 3: Data files where individual consecutive samples are aggregated into gaze events (fixations, saccades).	Eye tracker	.csv
"[subject_id]_[language]_[session_id]_reading_data_v00.csv"	Gaze event data	Stage 4: Reading data files where gaze events are mapped to individual words and then sorted sequentially (i.e., by the original word order of the presented text).	Eye tracker	.csv

The following table describes any additional data collected, e.g., the dataset contains demographic information or other relevant information:

Table 2: Additional data

Name	Data type	Description	Instrument	Format
"[subject_id]_[language]_[session_id]_exp_data_v00.csv"	Experiment data	Interest file consists of information about the experiment itself (i.e., interest area file with trial information, e.g., experimental condition, trial ID, item ID)	Computer	.csv
"[subject_id]_[language]_[session_id]_annot_stim_v00.csv"	Annotated stimulus data	the annotated version of used stimulus (incl. part of speech, lexical frequency)	Computer screen	.csv
"[subject_id]_[language]_[session_id]_demogr_data_v00.csv"	Demographic data	For each participant, demographic information about e.g., age, sex, socio-economic status, etc. were collected. The dataset contains only anonymized data.	Computer: Digital questionnaire	.csv
"[subject_id]_[language]_[session_id]_compreh_data_v00.csv"	Questionnaire data	Comprehension questions	Computer: Digital questionnaire	.csv
…	*Any other additional questionnaire data*	*Any other additional questionnaire which could be of any interest can be stated here.*	*Paper & Pencil* *or* *Computer: Digital questionnaire*	…

2. Ethics Approval

At large, every collection site within the MultiplEYE project should comply with the following principles and be ethically responsible within their research practice which is the respectful treatment of people who place themselves at the service of science for the purpose of research; principles (for an example, see ethical guidelines from the German Psychological Society, DGPs, from 2018):

- ✓ respect for autonomy

- ✓ non-maleficence
- ✓ beneficence
- ✓ justice

The responsible researcher at each data collection site must ensure the respectful treatment of their local ethics requirements. However, MultiplEYE provides support for every member of the COST Action to apply for ethics approval. We provide certain templates and support for the most common questions and answers (specific to the MultiplEYE experiment procedure) regarding ethics application.

3. Data Privacy & Protection

In general, MultiplEYE and all data collection sites are committed to ensuring the protection and security of all data that is collected and held. That is, MultiplEYE and all data collection sites are dedicated to the principles and guidelines inherent in the General Data Protection Regulation (GDPR), and particularly to the concepts of privacy by design, the right to be forgotten, and consent (Art. 25, Art. 17, Art. 4(11), and Art. 7 of the GDPR). In addition, we aim to ensure:

- transparency with respect to the use of data (processing must be traceable for the person concerned at any time)
- that any processing is lawful (by consent or legal permission), fair, transparent, and necessary for a specific purpose
- that data is accurate, kept up to date and removed when no longer needed
- that data is kept safely and securely.

MultiplEYE is generally responsible for the data protection concept and its systematic implementation. Moreover, the person(s) in charge at each data collection site is responsible for following data privacy regulations and complying with local data protection regulations and laws (if they differ from the GDPR regulations) during the collection process. After the collection process is done, the MultiplEYE is responsible to store and prepare all collected data for sharing. Secure storage will be provided by the University of Zurich (contact: Lena Jäger, jaeger@cl.uzh.ch) under the Swiss data protection law (Datenschutzgesetz, DSG).

Access to data during the collection process, including the processing of any personal data, is handled in accordance with the data protection regulations as well as ethical guidelines.

Therefore, all personal data (e.g., real names, demographics) must be immediately anonymized or pseudonymized and will not be a part of the eye-tracking corpus. Personal data such as contact data will be replaced with a unique identifier. This unique identifier (i.e., participant IDs) consists of a (random) numeric or alphanumeric ID, or pseudonymization code, which would not allow identification of the natural person. The participant's ID is used consistently throughout the study without revealing any personal information. Furthermore, we understand personal data as any information relating to a natural person that can be used to identify that natural person directly (by name, personal identification number, etc.) or indirectly (by IP address, any characteristics unique to the individual, etc.). Since MultiplEYE collects eye-tracking data and this data is considered a biometric characteristic and therefore personal data under GDPR, there is no concern regarding GDPR compliance since the identification of an individual through this data would require disproportionate effort.

This means that it would be difficult, impractical, or unreasonable to use eye-tracking data to identify an individual without investing a significant amount of time, resources, and expertise. Therefore, the risk of unauthorized identification or processing of personal data is low, and the GDPR does not apply in this case.

It is important to note, however, that even if GDPR does not apply, ethical considerations and best practices for data management should still be followed by all MultiplEYE members to ensure that participants' personal data is treated with respect and kept secure.

Apart from that, data collectors of MultiplEYE may only use contact data (e.g., real names and email addresses) as personal data in order to recruit participants for the eye-tracking experiment. All contact data must be deleted by the person(s) responsible for the data collection process at each collection site after it is no longer needed. When data collection / the reading experiment starts, the eye tracking data is collected along with a participant’s ID which is separated from any personal data such as contact data, so it is never linked to real names or similar.

In some cases, and this might only be applicable to a few collection sites depending on their research interests, personal data such as contact data (i.e., real names and email addresses) is linked to the participant's ID through corresponding tables or files, for example, when there

are more than one reading sessions planned with the same participants and they need to be contacted again. In this case, the correspondence tables or files which link each participant to a unique identifier, will then also be destroyed after a final data check as soon as the collection process is done.

Still, in order to prevent the identification of a person who participated in the eye-tracking experiment, it must never be possible to establish a link between contact data and the participant ID used for the experiment. Collected demographic data (e.g., sex, age, etc.) is only linked to the participant's ID and never to their real names or any other personal data which would allow an immediate identification of the natural person.

All (other) data relevant to the eye-tracking data corpus (e.g., raw and aggregated data, crucial metadata) contains only anonymous information. That data is documented by all MultiplEYE data collectors and (pre)processed in a standardized manner (i.e., all data are pre-processed by the MultiplEYE pre-processing team) in order to be shared for reuse. Consequently, access to the collected data (except for personal data) is primarily granted, besides to each collection site that collected the data, to all MultiplEYE leadership positions (i.e., Action Chair, Action Vice Chair, Grant Holder Scientific Representative, Grant Awarding Coordinator, Science Communication Coordinator, and Grant Holder Manager) as well as to all leaders of Working Group 1 (WG1: “Enabling eye-tracking data collection”) and the data preprocessing team of WG1.

Furthermore, the data are shared in the RDC at ZPID as well as at the repository “PsychArchives” under the German data protection law (*Datenschutzgrundverordnung*, DSGVO). For the MultiplEYEStore database, we select certain sharing levels that enable the provision of public use files, allowing open and immediate access to the data of MultiplEYE (which does not include any personal data), and therefore fostering its reuse and widespread dissemination. These sharing levels are applicable in all cases in which there are no plausible reasons (such as the inclusion of human subjects data or research ethics) for restricting access and use from the perspective of the scientific community (see also Section 9 of this DMP).

3.1. Informed consent

In the experiments involved in data collection for the MultiplEYE eye-tracking data corpus, all collected data is to be processed in a legally compliant manner on the basis of the informed consent of the participants. Part of the informed consent we are using is the clear declaration

of intent to participate in data collection, storage, processing, analysis, and sharing. Moreover, informed consent provides a clear and comprehensible explanation of the benefits and risks when participating in the experiment, and outlines the data collection process, the data processing purposes, the storage, and (re)use of data. A clear declaration of intent can be given as long as

a) the participant is capable of giving consent,

b) the participant consents actively and in an informed manner (i.e., by signing the informed consent).

Participants are informed before the collection of data and before the experiment starts. Depending on local regulations and laws at every collection site, the informed consent will be handed out in written or digital form. The signed consent forms (on paper or digital) are stored in a protected place, and for as long as the participants' data processing (based on the consent). If using paper consent, the forms are stored in a secure place, for example inside a lockable cabinet. If using digital consent forms, they are stored on any device (e.g., network drive) protected through encryption or password. The digital text of consent cannot be edited, and the full contents are unalterable, once they are signed.

4. Intellectual Property Rights & Copyright

In terms of intellectual property rights and copyright, MultiplEYE agrees on the following:

- All copyright and other intellectual property rights in any work developed during the MultiplEYE action will vest in us and all authors who contribute data to the eye-tracking data corpus. There will be licenses (e.g., Creative Commons licenses) that give everyone from individual authors or creators to large institutions a standardized way to grant the public permission to use their intellectual work under copyright law (see also Section 9 of this DMP).
- Moreover, shared primary data will not be subject to copyright concerns. Here, primary data means the recorded (raw) eye-tracking data in their first digital transmission (i.e., stage 0 and stage 1 of data processing), but otherwise in a completely unedited form.
- Stimuli (e.g., reading material) used for the eye-tracking experiments will include the usage of either copy-free texts or text excerpts, or (as in most cases) small excerpts from copyrighted texts and books which are considered as “fair use”. Since MultiplEYE

uses stimuli in various languages with different origins, for different collection sites and countries with (possibly) differing copyright laws, each data collection site is asked to check their local regulations and act accordingly.

- In general, we state and agree that all chosen copyrighted text excerpts and their translation must comply with the relevant copyright regulations of the countries of their origin. To name the most common and well-known ones, the European Union copyright directives (see Article 5 "Exceptions and limitations"), the U.S. American law, the so-called *fair use* doctrine (Section 107 of the U.S. Copyright Act), as well as with the right of *fair dealing* under UK law (Copyright, Designs, and Patents Act [CDPA] 1988, Section 29 & 30; see also exceptions to copyright by the Intellectual Property Office UK), and with the German copyright law ("Urheberrechtsgesetz [UrhG]", see esp. subdivision 1 & 4 of division 6 "limitations on copyright through uses permitted by law"). These state that limited and fair use of copyrighted material for research purposes (and others) is allowed when the copyrighted material is used by research authorities (i.e., non-commercial / government scientific institutions, research institutes, and universities) and when complying with a list of relevant factors regarding the purpose (i.e., non-commercial, education, research, etc.), the quantity (amount limited depending on purpose), (legal) acquisition, acknowledgment to creator(s) and so on. Furthermore, we agree that all relevant factors are to be considered when choosing the copyrighted stimuli for our research project.

  Since the UK, U.S. American, and German copyright laws differ in some definitions regarding limitations when using copyrighted material for research, we therefore mostly refer to the German copyright law (UrhG) as our baseline of this DMP.
- We created a list overviewing all 37 member countries of MultiplEYE (number of countries from May 2023) and their individual copyright regulations. The list can be found in Appendix 1. Within this list, the duration of copyright and the general copyright terms are shown - information that is crucial for the project's stimuli selection.
- Based on the German copyright law (and therefore in compliance with the EU copyright guidelines, and other copyright regulations from the previously mentioned list) we agree to the following:

- Regarding the purpose, MultiplEYE ensures to use of copyrighted excerpts lawfully (based on Section 60c and 60d UrhG) and only within the frame of our research (i.e., as stimuli for the reading experiment).
- We only choose copyrighted material that is lawfully accessible (which means that the material can be accessed from a library or acquired from a store). Copies of the materials are to be deleted when they are no longer needed (i.e., after the research project is completed) based on Section 44 of UrhG.
- Reproduction or copies of the used copyrighted texts are made available (e.g., online, on a cloud server, or similar) for MultiplEYE members only who are involved in the research (i.e., data collectors, WG 1, people who are monitoring the quality of the research). The making available to the public must be terminated as soon as the research or the monitoring of the quality of the research has been concluded.
- The amount of the excerpt used is limited to the extent needed “by the purpose” (Section 51 UrhG). Very short text excerpts from copyrighted books are used, which do not exceed 5% of the entire work (i.e., the total number of pages of a copyrighted book) as the UK copyright law (fair dealing) states. According to Section 60c of the UrhG, 15 % of a copyrighted text can be used for research purposes, when not published or distributed publicly (means, the original version or copies of the copyrighted material are not allowed to be published or to be used in any commercial way). Copyrighted (printed) work with a low scope (i.e., not more than 25 pages) can be used completely.
- The use of the copyrighted excerpts will not have any effect on the original work (see also Section 62 UrhG).
- We are obligated to always indicate the source and name(s) of the author(s) or creator(s) of the copyrighted source material (Section 62 UrhG). Text usage, translations, and any other relevant information (including metadata) are documented in a standardized manner (see also Section 8 for further explanation regarding data documentation).
- Furthermore, we assure that representations of the text or book excerpts (i.e., PDFs, copies of books) will not be made, published, or disseminated in any way.

No commercial gains will be made from them. Recording eye-tracking data that captures text from copyrighted material, it is essential to ensure that this data is used solely for your research purposes and not for publication or dissemination.

- ➢ The copy-free texts are official / institutional texts (e.g., “The Universal Declaration of Human Rights”), self-created texts (e.g., the MultiplEYE introduction text), publicly available texts with a license for public use (e.g., CC BY-SA 4.0), or very old texts for which the copyright has expired (i.e., 70 years after the death of the author). For all countries within the European Union (EU), copyright safeguards the intellectual property for a period of 70 years following the death of the creator. In the case of a jointly authored work, it applies for 70 years after the death of the last surviving author[1] (see also Sections 64 and 65 of UrhG). This “70-years rule” encompasses in most cases translations when they closely resemble or are identical to the original work. Essentially, copyright protection extends to translations only if they represent a translator's personal intellectual creation. This occurs when the translation displays a certain level of creativity distinct from the original (and this is a mostly subjective assessment from someone who knows both the original language and the translation language).
- ➢ MultiplEYE assures the use of translations that closely align with the original texts (e.g., translations of “The Emperor’s New Clothes”), ensuring that they faithfully reflect the intended meaning. By adhering to this practice, we maintain compliance with the "70-years rule", which continues to apply.

- MultiplEYE provides a list of copy-free and copyrighted fair-use text excerpts in different languages which are available for all data collection sites to allow a standardized procedure and comparable research outcomes. In some cases, texts must be translated by MultiplEYE members (if a translated version does not exist publicly), so that they can be used at certain locations/countries.

[1] Outside of the EU, in countries that are signatories to the Berne Convention, the duration of copyright protection may differ, but it generally lasts for a minimum of 50 years after the author's death.

Table 3: Stimuli final selection and associated copyright

Name / Text ID	Title	Author(s)	Original language	Copyright Original	Copyright Description	Source
PopSci_MultiplEYE	Welcome to MultiplEYE	Nora Hollenstein, Ana Matić Škorić, Lyndsey Denton-Fray	en	no copyright	self-created text, copyright vest in us	
Ins_HumanRights	Universal Declaration of Human Rights	UN General Assembly	en	no copyright	Official text, public usage allowed	https://www.un.org/en/about-us/universal-declaration-of-human-rights
Ins_EURLex	REPORT FROM THE COMMISSION TO THE COUNCIL: Progress report on a Learning Mobility Benchmark	European Commission	NA	no copyright	Official text, public usage allowed, under license CC0	https://eur-lex.europa.eu/content/legal-notice/legal-notice.html
Lit_Alchemist	The Alchemist - Chapter 1	Paolo Coelho	pt	copyrighted - but research usage permitted	The Original by Paulo Coelho and the English translation by Alan R. Clarke is copyrighted under International and Pan-American Copyright Convention. The Article XII of the Pan-American Copyright	https://scholarlycommons.law.wlu.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=415

					Convention says: the reproduction of brief extracts of protected works for the purposes of instruction, research or criticism is permitted.	8&context=wlulr
Lit_EmperorClothes	The Emperor's New Clothes	Hans Christian Andersen	da	no copyright	"70-years rule" applies: creator of original is more than 70 years dead, so copyright expired	https://forskerportalen.dk/en/research-and-copyright-admin/
Lit_HarryPotter	Harry Potter and the Philosopher's Stone - Chapter 1: The boy who lived	Joanne K. Rowling	en	copyrighted - but research usage permitted	The right of fair dealing under UK law applies here (Copyright, Designs and Patents Act [CDPA] 1988, Section 29 & 30; see also exceptions to copyright by the Intellectual Property Office UK). Which means excerpts of copyrighted material can be used for (non-commercial) research purposes.	https://www.gov.uk/guidance/exceptions-to-copyright
Lit_MagicMountain	The Magic Mountain Foreword	Thomas Mann	de	copyrighted - but research usage permitted	German copyright law applies here: 1 & 4 of division 6 "limitations on copyright through uses permitted by law". These state	https://www.gesetze-im-internet.de/englisch_urhg/e

					that a limited and fair use of copyrighted material for research purposes (and others) is allowed when the copyrighted material is used by research authorities (i.e., non-commercial / government scientific institutions, research institutes, universities).	nglisch_urhg.html#p0321
Lit_NorthWind	The North Wind and the Sun	Aesop	NA / grc	no copyright	ancient greek literature: “70-years rule” applies - copyright is expired	https://web.archive.org/web/20090319233331/http://portal.unesco.org/culture/en/files/37873/12221737361212193_english%5B1%5D.pdf/2121_93_english%5B1%5D.pdf
Lit_Solaris	Solaris - Chapter 2: The Solarists	Stanisław Lem	pl	copyrighted - but research usage permitted	ACT of 4 February 1994 ON COPYRIGHT AND RELATED RIGHTS “Article 27. Research and educational institutions shall be allowed, for	https://en.wikisource.org/wiki/Polish_Copyright_Law

					teaching purposes or in order to conduct their own research, to use published works in original and in translation, and to make copies of fragments from the disseminated work for the same purpose.	
Lit_BrokenApril	Broken April - Chapter 3	Ismail Kadare	sq	copyrighted - but research usage permitted	Scientific use allowed for copyrighted material. Reproduction of a published work without the author’s approval and without payment or remuneration, according to the laws, is allowed only for personal use (use for research and scientific purposes included).	https://archive.ph/20071030185836/http://www.wipo.int/clea/docs_new/en/al/al001en.html
Arg_PISARapaNui	Rapa Nui	OECD	en/fr	no copyright	PISA text, public usage allowed, **they** are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) license.	https://www.oecd.org/termsandconditions/
Arg_PISACowsMilk	Lehma piim	OECD	en/fr	no copyright	PISA text, public usage allowed, **they** are licensed under the Creative Commons	https://www.oecd.org/termsandconditions/

					Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) license.	
Enc_WikiMoon	Moon	Wikipedia	en	no copyright	Wikipedia text, Text is available under the Creative Commons Attribution-ShareAlike License 4.0	https://en.wikipedia.org/wiki/Moon

- There are two practice texts which are used for practice trials before the actual experiment starts. The collected data of the practice trials will not be shared / will not be a part of the MultiplEYE data collection.

5. Data Storage, Backup & Extermination

All data created at all stages, which is shown in the data description tables, must be stored in accordance with the following agreements:

- At all times, data storage and backup are done in accordance with the data protection guidelines which have already been stated above.
- All data collectors of MultiplEYE at each collection site are responsible for their own data storage and backup during the data collection process. This includes storage of all data collected for the eye tracking data corpus as well as any personal data (i.e., demographic data and/or contact data: real names, email addresses). Therefore, the individual institution or university at each collection site provides a network, drive, or cloud system with appropriate space as well as proper access rights control (control of who can see and edit folders and files on the network drive) or with proper encryption to ensure data security. Storage of any personal data on unencrypted portable external drives or on unencrypted USB storage devices is prohibited. Personal data such as contact data or, if applicable, a corresponding table identifying participants by linking participant IDs with contact details (e.g., real names) are stored within an encrypted folder (for example by using 7-Zip encryption).

- Each collection site is advised to provide detailed documentation of their data storage to avoid data loss or to be able to trace certain work processes. This includes a data inventory list, backup list, and extermination list.

Example of data inventory list:

File name	Content description	Type	Version	Format	Soft-ware	Volume	Data protec tion	Access cate-gory	Responsib le person (name + affiliation)	Project end	Source/ storage location
20230304_MultiplEYE_natread_recfile_v01.asc	Recorded files eyetracker	Data matrix	V01	.asc	xLabs (version 2.6.1.)	~ 10 MB	yes	Restricted (data collectors only)		Archiving and sharing	Own server (+backup institution server)
…											
…											

Example of data backup list:

File name	Content description	Type	Version	Format	Soft-ware	Volume	Data protec tion	Access cate-gory	Responsib le person (name + affiliation)	Project end	Source/ storage location
20230304_MultiplEYE_natread_recfile_v01.asc	Recorded files eyetracker	Data matrix	V01	.asc	xLabs (version 2.6.1.)	~ 10 MB	yes	Restricted (data collectors only)		Archiving and sharing	Own server (+backup institution server)
…											
…											

Example of data extermination list:

File name	Content description	Storage location	Extermination due date (bound by contract)	Deleted on (date)	Deleted through (name of person responsible)	Comments
20230304_MultiplEYE_natread_contactdata_v00.txt	Contact data participants for recruitment	Own server (+ backup institution server)	Immediately	…	…	Immediately, when no longer needed
…			09/2026			Extermination at end of project
…						

- In general, all data collected and chosen to be used for the eye-tracking data corpus, is also stored and secured by the Action’s work group 1 data preprocessing team which includes the storages at a secure cloud service of the University of Zurich (only anonymized data).

- As previously mentioned, and according to the declarations of consent, personal data, such as contact information (i.e., real names and email addresses), which might also only be applicable for some collection sites, is only stored for as long as it is necessary for the further course of the project (for example, if there are more than one reading sessions with the same participants). In this case, the correspondence table which enables linking each participant to a unique identifier, will be destroyed after a final data check when the collection is finished.
- At large, no longer needed personal files are destroyed immediately.
- The signed declarations of consent (paper documents) are stored in a protected place (e.g., inside a lockable cabinet). If the consent forms are digital, they are stored and protected (through encryption / providing a password) on any device (e.g., a network drive). The text of consent forms must be locked to prevent editing, and the full contents unalterable, once they are signed.
- The signed consent forms (on paper or digital) are stored for as long as the participants' data processing (based on the consent). The retention period will depend on national laws and may vary depending on the country of each collection site. Generally, however, consent forms should be kept for at least as long as the data from the study itself.

6. Data Organization

Generally, MultiplEYE recommends that all collection sites keep a clear directory organization and structure of all collected and documented data to make it easier to locate files and versions. When collecting/receiving the data from all collection sites, MultiplEYE uses this standardized naming convention throughout the project. MultiplEYE advises all data collection sites to adhere to a well-organized structure to support the MultiplEYE team in managing the data corpus.

6.1. Folder structure

Files should be structured in a hierarchical order from a general, high-level folder to more specific folders nested inside. The hierarchy of folders should be consistent and logical. We provide guidelines and recommendations on how to structure folders and files specifically for this project and make them available for all data collectors of this COST Action.

The guidelines can be found here. We propose such a structure because standardized and well-structured storage of the data contributes to efficient collaboration. In addition, adhering to a consistent structure makes it easier to keep an overview of the completeness of the data in order to ensure the data’s quality and enable its reusability.

6.2. Folder and file naming conventions

As a part of the COST Action Project “MultiplEYE”, the **titles or folders** should be named accordingly and should be used consistently. This should include the terms “MutliplEYE_[the tested language: use ISO-639-1 language codes]_[name/short form of lab_country of data collection: use ISO-3166 country codes]”. There is also the option to add the used method (self-paced reading [=spr] vs. eye tracking [=eyetr]) and session number (if more than 1 session took place with the same participants).

Files should be named in accordance with the following rules:

- use only underscores to combine elements in the file name (_)
- file names should include

  1) the participant or subject ID,

  2) the tested language (use ISO code),

  3) (if applicable) the session ID/number,

  4) method used (self-paced reading or eye-tracking),

  5) type of data (i.e., raw data, reading data, questionnaire data)

  6) the version (see also section about file versioning)

  7) presentation of file format extension (e.g., .doc, .xlx, .csv)
- If there is the need to add dates to file names, use ISO 8601 as a standard.
- Example: "PD023_DE_s02_eyetr_reading_data_v00.csv"

6.3. Versioning

MultiplEYE agrees on using the following standard of versioning:

- using at the end of every file name (but before the format extension is being presented) a “v” for version and add the version number
- version numbers start at 00, continuing with 01, 02, 03, 04, …09, 10 and so on
- also, initials of the person who created this version must be added, use small letters only (e.g., “mm” for the person named Max Mustermann)

7. Data Quality

In order to ensure high data quality for the MultiplEYE data collection, data must be consistent across participants and languages. Furthermore, we ensure that the collected data is reliable with consistent as well as repeatable measures and results. The following data quality criteria apply to the entire data collection of MultiplEYE and are obligated to be observed on all data collection sites. Certain data quality checks are built into the data preprocessing and are performed during the different phases of the process.

- **Data Consistency**: All data goes through the preprocessing pipeline of the MultiplEYE preprocessing team! Data is preprocessed using a preprocessing pipeline containing 4 preprocessing steps: 1) Converting non-human readable recorded raw data files (i.e., .edf-files) into human-readable formats (i.e., .asc-files), 2) parsing raw data files, 3) detecting gaze events, and 4) compute reading measures. All preprocessing steps will be done by using the Python software package *pymovements*[2].
- **Data Reliability**: We ensure that the eye-tracking data are reliable, with consistent and repeatable results (e.g., through sanity checks).
- **Data Validity**: We ensure that the eye-tracking data are valid, with accurate and meaningful measurements of eye movements (we use sanity checks for checking for missing data, outliers, incorrect formatting, or other types of errors that can affect the accuracy and reliability of the data).
- **Data Accuracy / Calibration Accuracy**: We ensure that the eye tracker has been accurately calibrated before and also within each data collection session (that is, every time before each new trial [=text]). We calibrate and validate (1) at the start of the experiment. Validation is performed immediately after calibration, and additionally (2) before each new text (i.e., 13 times within a session), a validation check is also performed (3) at the end of the recording.
- **Data Completeness**: We ensure that all required data are collected for each participant and that there are no missing or incomplete datasets to a certain amount; complete datasets are to be prioritized for reuse, minimum completeness: ~10% of the experiment completed (i.e., minimum 1 reading task / 1 text completed)

[2] https://github.com/aeye-lab/pymovements

- **Precision** of fixations: We check the proportion of fixations which are outside of the area of interests (AOIs), the proportion of fixation which is not on the first word of the first line, the proportion of the fixation “line” which is not horizontal (a line that runs obliquely up or down, i.e., deviates from the horizontal line), the proportion of line jumps, the proportion of events when many consecutive words have not been read.
- We provide certain **data quality checks and measures within data pre-processing** (i.e., data check of each session).

Other relevant quality criteria:

- **Participant Quality**: We ensure that participants are suitable for the study, and have normal or corrected-to-normal vision. Ensure that participants are suitable for the study, and have normal or corrected-to-normal vision. The participants’ recruitment and invitation will include those inclusion criteria as a requirement to participate in the experiment. Other questionnaires and tests before participation will further ensure the participants’ quality.
- **Task Quality**: We ensure that the task being performed by participants is appropriate and engaging and that it is well-defined and standardized across participants and languages. We have tested the length and reading times of texts, made sure that the choice of texts is appropriate, and that the presentation on screen is of high quality.
- **Lab Setup**: We provide standardized instructions and guidelines on how to set up the lab and experiment to ensure high-quality standards and to avoid common mistakes.

Datasets that meet certain quality criteria are considered worth sharing and will be long-term preserved by the repository PsychArchives.

In general, MultiplEYE agrees on including all kinds of quality levels. This means, in some cases, data quality could differ due to the use of different hardware (i.e., eye tracker, low-cost / low-frequency eye tracker). Regions with fewer resources might be more affected by this than others, but they are still of much interest to us because of their rare language data. This, of course, requires an open & transparent data documentation process, as well as transparent data quality reviewing, so that data (re)users are aware of the data’s quality level and can decide for themselves how to (re)use it.

For reviewing purposes and to avoid data errors or poor data quality caused, for example, by incorrect resolution or code error, as well as to check if the data is processable, each collection site is asked to send some samples of their eye-tracking data to the data preprocessing team and data quality reviewer team of Working Group 1 before data collection starts (i.e., when test runs are made).

8. Data Documentation

All members of MultiplEYE involved in data collection, which means the person(s) responsible for the data collection process at each collection site, provide full documentation of their research data to be in compliance with the FAIR principles. All documentation data will be shared along with the eye tracking datasets through the research data center (RDC) of the Leibniz Institute for Psychology (ZPID) and via an open-source repository PsychArchives.

The following documentation steps are included in a full research data documentation and must be included when sharing all collected data:

1) **Metadata documentation**: This contains all relevant metadata that describes the study or research background and is needed to interpret the actual data. This documentation provides primarily descriptive (e.g., title, authors, etc.) and content metadata (e.g., research objective, study design, research instruments, etc.).

   For reasons of consistency, MultiplEYE provides a standardized metadata form for a detailed study description and is made available on the project website to all data collectors as a “form to fill out” to prepare their data for sharing (it will be shared as a PDF file along with the eye movement data when reusers access a dataset). This metadata form is based on the Dublin Core metadata standard and aligns with the metadata scheme of the research data center where the MultiplEYE data will be stored.

2) **Codebook documentation**: MultiplEYE requests from all collection sites to include an extended codebook within their research data documentation. The codebook provides detailed variable descriptions and important metadata to enable reusability and also findability of the shared research data.

   We also advise all data collectors who are responsible for the data documentation to export the codebook in a machine-readable format (preferably .json). In other cases, the codebook is provided as a .csv file.

Content of codebook

Attribute	Description
Variable name	A short, meaningful name of the variable, relating to the variable's content; only small letters should be used, spaces should be replaced by an underline (e.g., us_pos)
Variable label	A more comprehensive description of the variable (e.g., unconditioned stimulus, positive valence)
Item text	The exact question of a questionnaire / instruction text
Value labels	Describes the value range of a given variable, in case of paradata (e.g., instructions, material) the relevant file name wherein the information should be indicated
Missing	Specifies missing values (e.g., -99 for omitted questions)
Measure	The measure used for the dependent variable as defined in the study documentation

3) **Data preprocessing documentation**: Here, the steps from raw data to analyzed data are provided (i.e., all steps of the pre-processing pipeline, which are: 1. converting recorded files into human-readable formats, 2. converting data formats into csv files with one row per recorded sample, 3. event detection, 4. computing reading measures (gaze duration, re-reading time, regression probability, etc.). The preprocessing team of WG1 is responsible for adequate documentation of those steps. They provide comprehensive details on not only the software packages that were used to make these transformations but also a plain summary of how the data was altered when using these procedures (i.e., gaze detection). Here it is important to mention, that the preprocessing team uses the software *pymovements*[3], which is an open-source Python package for processing eye movement data. Source code is documented on GitHub and is made publicly available (see also next section about "Data Sharing & Licensing'').
4) **Other documentation**:

[3] https://github.com/aeye-lab/pymovements

- All data collectors should provide a session documentation sheet that records any unexpected events, mistakes, or errors. This documentation should be done during a session (real-time) and for each session. The sheet should include trial IDs and text IDs, and the errors that might have happened during the session (e.g., abortion of a trial or the participant was coughing/talking/interrupting the task within a trial). This has to be documented immediately (on paper) and later digitized.
- Also, a deviation documentation should be provided for each data collection, in case any deviations within the experiment presentation, the lab setup, or the stimulus selection (and many more) occurred. Any deviations from our standardized procedure should be recorded within this form.

9. Data Sharing & Licensing

MultiplEYE shares all collected and documented (and anonymized) data via the Research Data Center (RDC) and the open-source repository PsychArchives, both belonging to the Leibniz Institute for Psychology (ZPID) in Germany. The website of the RDC provides a platform for MultiplEYE (our own database called “MultiplEYEStore”) so that the data corpus can be found, is accessible, and is searchable.

The data will be shared and identified by a persistent identifier (PsychArchives uses Digital Object Identifiers, “DOI"), so the data can be cited easily and uniquely. Since PsychArchives is a long-term archival system, all data belonging to the MultiplEYE data corpus (i.e., eye-tracking data at all stages of the preprocessing pipeline, from raw to aggregated data, as well as all relevant metadata, codebooks, and additional study documents) will be long-term preserved. The RDC and PsychArchives provide different sharing levels which control the access of all data. MultiplEYE uses sharing level 0 (public use file).

Moreover, code and software packages for data pre-processing are made publicly available on open-source repositories as well. Source code and documentation are shared via GitHub. A Python library for the preprocessing pipeline and a Wiki for available dataset resources are already outcomes of MultiplEYE and are shared here:

https://github.com/aeye-lab/pymovements

https://github.com/norahollenstein/cognitiveNLP-dataCollection/wiki

All COST Action leadership positions are granted to publish the data in journals with the consent of those who collected it. Also, local data collectors are to be mentioned/cited when others (re)use their data.

Therefore, MultiplEYE agrees to use the following licenses for granting usage rights to anyone interested in using our work in accordance with the FAIR principles. Furthermore, we share all data relevant to the eye-tracking data corpus under the Creative Commons licenses which are considered genuine Open Access licenses. Specifically, we choose the license CC-BY-SA 4.0 which allows the distribution, reproduction, modification, or use in any way, as long as the authors/creators are mentioned as originators. In this way, the CC-BY license helps achieve greater visibility for academic publications and their authors/creators. Furthermore, CC-BY-SA 4.0 binds users of a work to the original license. The SA (Share Alike) module is a useful alternative for authors/creators who want to prevent others from making money with their content, as it requires "sharing under the same (license) conditions, without limiting the reuse potential of the data.

Since Creative Commons is not recommended for licensing software, the *pymovements* package and the MultiplEYE experiment code *are* licensed under the open source license "MIT License" which gives express permission for users to reuse code for any purpose (modify, change, etc.) as long as users include the original copy of the MIT license in their distribution.

10. Statement of Agreement

All members of the COST Action MultiplEYE agree to carry out data management in accordance with the project's data management plan.

11. Acknowledgements

We thank all members of the MultiplEYE COST Action who are contributing to the MultiplEYE data collection and who provided helpful comments to this document.

This work was partially funded by *swissuniversities* through a CHORD-B Swiss Open Research Data Grant and is supported by COST Action MultiplEYE, CA21131, funded by COST (European Cooperation in Science and Technology).

References

Abele-Brehm, A., Gollwitzer, M., Steinberg, U. & Schönbrodt, F. (2019). Attitudes towards Open Science and public data sharing: A survey among members of the German Psychological Society. *Social Psychology*, *50*, 252 – 260. https://doi.org/10.1027/1864-9335/a000384

Deutsche Gesellschaft für Psychologie DGPs (Eds.) (2018). Ethisches Handeln in der psychologischen Forschung [Ethical behavior in psychological research]. Göttingen: Hogrefe.

Krakowczyk, D. G., Reich, D., R. Chwastek, J., Süss, A., Prasse, P., Jakobi, D. N., Turuta, O., Kasprowski, P., Jäger, L. A. (2023). pymovements: A Python package for eye movement data processing**.** *ETRA 2023*, 10.1145/3588015.3590134

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Comput Biol* 11(10): e1004525. doi:10.1371/journal. pcbi.1004525

Mikser, T., Simms, S., Mietchen, D., Jones, S. (2019). Ten principles for machine actionable data management plans. *PLoS Comput Biol* 15(3): e1006750. https://doi.org/10.1371/journal. pcbi.1006750

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. I., Appleton G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18