easier

intElligent Automatic
SIgn languagE tRanslation

# D11.2 DATA MANAGEMENT PLAN V1

Revision: v.1.0

| | |
|---|---|
| **Work package** | WP 11 |
| **Task** | Task 11.3 |
| **Due date** | 30/06/2021 |
| **Submission date** | 15/07/2021 |
| **Deliverable lead** | Martel |
| **Version** | 1.0 |
| **Authors** | Giacomo Inches (Martel) leading, contributions from all WP leaders |
| **Reviewers** | Köberlein Michael (NURO), Felix Burkhardt (AUD) |

| | |
|---|---|
| **Abstract** | This document identifies the best practices and specific standards for the generated data, assess their suitability for sharing/reuse in accordance with official EC guidelines. The Data Management Plan will be prepared, detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use and how it will be curated and preserved – adhering to the FAIR guiding principles as adopted by the EC for research data management.<br>A first version of the DMP will be delivered within the first six months of the project. Updated versions of the DMP will be provided with the interim and periodic reports as relevant. |
| **Keywords** | Data Management Plan, Data Protection, Open Data, Open Research Data Pilot |

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

## Document Revision History

| Version | Date | Description of change | List of contributor(s) |
|---|---|---|---|
| V0.1 | 03/06/2021 | 1st version of the ToC, partners contribution requested | Martel |
| V0.2 | 11/06/2021 | WP1, WP3 WP4, WP7, WP10, WP11 Contributions | UZH, UNIS, Martel, EUD |
| V0.3 | 21/06/2021 | WP2, WP5, WP9 contributions | ATHENA, RU, CNRS |
| V0.4 | 29/06/2021 | WP6, WP8 contributions | UHH, NURO |
| V0.5 | 08/07/2021 | Consolidated version ready for internal review | Martel |
| V0.6 | 13/07/2021 | Input from Reviewer 1 and Technical Leader | AUD, ATHENA, UHH |
| V0.7 | 14/07/2021 | Input from Reviewer 2 | NURO |
| V1.0 | 15/07/2021 | Final check and ready for submission | Martel |

## DISCLAIMER

The information, documentation and figures available in this deliverable are written by the "Intelligent Automatic Sign Language Translation" (EASIER) project’s consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

## COPYRIGHT NOTICE

| | | |
|---|---|---|
| **Project co-funded by the European Commission in the H2020 Programme** | | |
| **Nature of the deliverable:** | **R** | |
| **Dissemination Level** | | |
| **PU** | Public, fully open, e.g. web | ✓ |
| **CL** | Classified, information as referred to in Commission Decision 2001/844/EC | |
| **CO** | Confidential to EASIER project and Commission Services | |

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

# EXECUTIVE SUMMARY

This deliverable is the first version (V1) of the EASIER Data Management Plan (DMP), defined at the end of M06. DMP is a living document that identifies which data is produced during the EASIER project, who owns that data, how it will be documented, how it will be preserved, and with whom and under what form it will be shared. Two additional releases of DMP are foreseen at M18 (V2) and M36 (V3).

Most of the data produced during the project will correspond to documentation, source-code and experimental results. Additional data containing personal information will be anonymized and will comply with to the General Data Protection Regulation (GDPR) and the highest Ethical standards.

The purpose of the DMP is to contribute to good data handling through indicating what research data the project expects to generate and describe which parts of the data that can be shared with the public. Moreover, it gives instructions on naming conventions, metadata structure, storing of the research data and how to make public data available.

This plan describes:

- what kind of data will be generated, collected, processed and shared,
- which methodology and standards will be applied during data collection and handling,
- the procedures for sharing and open access to the EASIER data and for curation and preservation of the data
- the procedures to comply with GDPR to ensure the protection of the involved companies' data, information and privacy rights.

As part of Horizon 2020, the EASIER project participates in the Pilot on Open Research Data (ORDP) a pilot action that aims to improve and maximise access to and re-use of research data generated respecting also the Findable, Accessible, Interoperable and reusable (FAIR) guidelines of the H2020 programme, which state that data will be made as available as possible, so long that does not negatively affect the commercial advantage of the partners.

# TABLE OF CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

# LIST OF FIGURES

Funded by the Horizon 2020
Framework Programme of the European Union

# LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

# ABBREVIATIONS

| | |
|---|---|
| **BSL** | British Sign Language |
| **CA** | Consortium Agreement |
| **DGS** | German Sign Language |
| **DMP** | Data Management Plan |
| **DoA** | Description of Action |
| **DOI** | Digital Object Identifier |
| **DSGS** | Swiss German Sign Language |
| **EMSL** | European Meta Sign Language |
| **FAIR** | Findable, Accessible, Interoperable and Reusable guidelines |
| **GDPR** | General Data Protection Regulation |
| **IP** | Internet Protocol |
| **LIS-CH** | Italian Sign Language of Switzerland |
| **LREC** | Language Resources and Evaluation (International Conference on) |
| **LSF-CH** | French Sign Language of Switzerland |
| **MT** | Machine Translation |
| **ORDP** | Open Research Data Pilot |
| **SL** | Sign Language |
| **TCP** | Transmission Control Protocol |
| **Vx** | Version x of the document e.g. v1, v2, v3, … |
| **WP** | Work Package |

Funded by the Horizon 2020 Framework Programme of the European Union

# 1 INTRODUCTION

## 1.1 ORDP PARTICIPATION

This deliverable D11.2 is of the type Open Research Data Pilot (ORDP), which by definition of the European Commission enables open access and reuse of research data generated by Horizon 2020 projects. The ORDP is constituted by two main pillars:

a. the Data Management Plan (DMP) and

b. an open access to research data.

A project that opts-in ORDP have to adhere to the following conditions:

- Develop (and keep up-to-date) DMP,
- Deposit the data in a research data repository.
- Ensure third parties can freely access, mine, exploit, reproduce and disseminate this data.
- Provide related information and identify (or provide) the tools needed to use the raw data to validate the research.

The ORDP applies to:

- The data (and metadata) needed to validate results in scientific publications.
- Other curated and/or raw data (and metadata) that are specified in the DMP.

Depending on the current consensus within the consortium some of the EASIER Artefacts might not be publicly available to comply to GDPR and preserve privacy of the individuals involved.

## 1.2 DATA MANAGEMENT PLAN

This document constitutes the first version of the Data Management Plan required by the ORDP nature of the deliverable and it is outlining how research data will be handled during a research project and after it is completed. It includes clear descriptions and rationale for the access regimes that are foreseen for all the collected data sets within the EASIER project.

The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project, therefore it describes the data management life cycle for all data sets that will be collected, processed or generated by the project. It will outline how research data will be handled during the research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved.

This first version of DMP is delivered within the first 6 months of the project while more elaborated versions of the DMP will be delivered at later stages of the project i.e. the second version of the DMP at M18 and the third one at the end of the project at M36. The DMP will need to be updated to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start.

Funded by the Horizon 2020
Framework Programme of the European Union

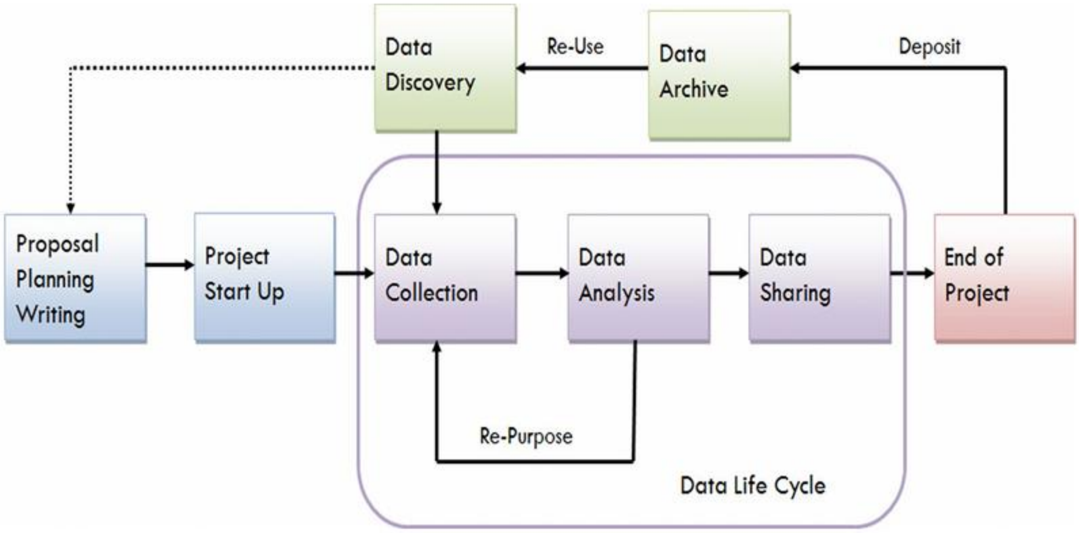

*FIGURE 1: DATA MANAGEMENT LIFECYCLE*

This first version of the DMP focuses on the identification of the different kind of data that each Work Package (WP) is or will be handling along the course of the project. The next versions of the DMP will complement this list and eventually enrich the properties each dataset has currently specified, as well as their metadata and tools for processing and repository location. Data are therefore organized on a WP level and presented in an exhaustive way in Section 2. In there we also describe the initial properties of each collection towards the FAIR (Findable, Accessible, Interoperable, Reusable) principles[1].

## 1.3 PUBLISHING INFRASTRUCTURE FOR OPEN ACCESS

While access to data is always guaranteed within the project partners and regulated by the Consortium Agreement (CA) and Description of Action (DoA) in the Grant Agreement, not all data generated within the project can be publicly released, due to privacy and confidentiality reasons e.g. full transcript of user interviews in WP1.

The data necessary to successfully complete the project Work Packages (WPs) will be shared without any restrictions amongst the WP partners either via internal repositories or direct communication. Public data will be made available at the project's website or other repositories, as appropriate. Users will be made aware of this data primarily through research publications, patent applications, dissemination activities, invited talks, social networks and the project website. Data will be made available to the project consortium as soon as it is available; to standardization bodies when required; and to the public at the due date of the derivable, and, in case a research publication is based on that, as soon as the paper is submitted (if submission is anonymous, this will be postponed). If access to confidential data is necessary by the public, restrictive measures will be put in place.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Funded by the Horizon 2020 Framework Programme of the European Union

Updates to these policies and link to repositories will be provided in the next versions of the DMP.

## 1.4 UNIFIED DOI STRATEGY

In accordance with the FAIR principles, any output of EASIER should be assigned a persistent identifier. Where the assignment of persistent identifiers is not part of the submission procedures (as is often the case when submitting papers to a journal), the partners involved in the creation of the output will take care of obtaining the persistent identifier(s) unless there are legal reasons not to do so. Where possible, there should be two identifiers, one for the product in general and one for the specific version to be released. If one of partners involved in creating the output has internal or by-contract access to a persistent identifier issuer, that partner will be asked to request the persistent identifier on behalf of the group of authors from EASIER. Where possible or even necessary, this involves submitting the output to the partner's long-term archive. Otherwise, the coordinating partner will choose a service such as ZENODO to submit the output to the long-term archive and obtain persistent identifiers.

When submitting the output to an archive, regulations forbidding the output to also be deposited elsewhere should not be accepted. When there is a choice, partners are asked to prefer DOIs over other persistent identifier schemes.

Funded by the Horizon 2020
Framework Programme of the European Union

# 2 DATA

In this Section we illustrate the data generated within EASIER, organized by WP. This division allows for better linking the dataset with their respective part of work. Another important remark to make concern the nature of the data and their purpose. Some of the dataset illustrated below are research dataset that will be include as part of the ORDP (e.g. multimedia files, annotations, source code), while others are data functional to the project management or dissemination and therefore not part of the ORDP.

## 2.1 IDENTIFIERS

Each data[set] is identified with the following format:

**EASIER_<WP number>_<serial number of dataset>_<data type>_<dataset name>**

The <*WP number*> correspond to those in the DoA, the <*serial number of dataset*> is a progressive number to uniquely identify the dataset within the same WP while the <*data type*> is mapped in the table below (Table 1). The <*dataset name*> is the arbitrary name of the dataset assigned by each WP leader.

*TABLE 1 IDENTIFIER TYPES OF DATA*

| Acronym | Data Type |
|---|---|
| SRC | Source Code |
| MUL | Multimedia content (audio, video, text) |
| PER | Personal data (email addresses, other personal information) |
| OTH | Other type of data (to be specified in each table below) |

## 2.2 PROPERTIES

In the following Sections, all the datasets are described with the help of the following properties:

- Identifier
- Dataset Summary
- Intellectual Property Rights
- Findability
- Accessibility
- Interoperability
- Security
- Ethical aspects
- Other issues

In the APPENDIX B we include the template shared among the WP leaders to collect the datasets information where these properties are explained into more details.

Funded by the Horizon 2020 Framework Programme of the European Union

## 2.3 DATASETS

The following table summarises the identified datasets and their scientific or administrative nature.

| Dataset Identifier | Nature |
|---|---|
| EASIER_WP1_01_PER_FocusGroups | Scientific |
| EASIER_WP3_01_SRC_EMSL | Scientific |
| EASIER_WP4_01_MUL_DSGS_broadcast | Scientific |
| EASIER_WP4_02_MUL_LSF-CH_broadcast | Scientific |
| EASIER_WP4_03_MUL_LIS-CH_broadcast | Scientific |
| EASIER_WP4_04_MUL_DGS_broadcast | Scientific |
| EASIER_WP4_05_MUL_BSL_broadcast | Scientific |
| EASIER_WP4_06_MUL_training_data_quality_estimation_models | Scientific |
| EASIER_WP5_01_OTH_postEditedData | Scientific |
| EASIER_WP5_02_MUL_verbalisingDiagramCorpus | Scientific |
| EASIER_WP5_03_MUL_qualityEstimationTrainingData | Scientific |
| EASIER_WP6_01_MUL_LREC-Anthology | Scientific |
| EASIER_WP6_02 MUL_Data_Collection_Tasks | Scientific |
| EASIER_WP6_03_MUL_Corpora | Scientific |
| EASIER_WP6_04_MUL_Lexical_Resources | Scientific |
| EASIER_WP6_05_MUL_Interlingual_Index | Scientific |
| EASIER_WP6_06 SRC_Data_Harmonization_Tools | Scientific |
| EASIER_WP7_01_MUL_annotated_video_data | Scientific |
| EASIER_WP8_01_MUL_VideoRecording | Scientific |
| EASIER_WP8_02_MUL_AudioRecording | Scientific |
| EASIER_WP8_03_PER_UserDataset | Scientific |
| EASIER_WP9_01_PER_ContactList | Scientific |
| EASIER_WP10_01_PER_NEWSLETTER | Administrative |

| | |
|---|---|
| EASIER_WP10_02_PER_WEBSITE | Administrative |
| EASIER_WP11_01_PER_Mattermost | Administrative |
| EASIER_WP11_02_PER_Mailman | Administrative |

## 2.4 DATA GENERATED IN WP1

In WP1, it was planned to collect crucial information to feed the reflections and the understanding of user needs and practices. Thereafter, to set up performance metrics (components and overall EASIER system) and develop recruitment strategies for evaluation studies. To achieve WP1's objectives, data are needed.

*TABLE 2 EASIER_WP1_01_PER_FOCUSGROUPS*

| **Identifier** | **EASIER_WP1_01_PER_FocusGroups** |
|---|---|
| **Dataset Summary** | **Responsible partner:** EUD, INT<br>**Purpose**: Data collection during focus groups to feed T1.1 (Analysis of user needs) and this will help to achieve the user centred design of EASIER mobile app and end-user tools.<br>The data collected will be used by WP2, WP4, WP5, WP7 and WP8 as the user involvement is the key to achieve EASIER's user acceptance.<br>**Type/format**: Transcripts anonymised<br>**Re-use of existing data**: N/A<br>**Data origin**: From EUD/INT focus groups<br>**Expected size**: N/A<br>**Data utility**: To all EASIER partners who wants to assess/understands user needs and practices, user expectation, etc. |
| **Intellectual Property Rights** | **Owner:** EUD/INT/EASIER<br>**Licensing:** N/A<br>**Dependency:** Only restriction is the use in the framework of EASIER project only<br>**Constraints:** N/A<br>**Timeframe:** Data only in the lifetime of EASIER project |
| **Findability** | **DOI:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Is data discoverable:** No<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** Only for EASIER partners and use to achieve EASIER's objective as the interviewed users agreed that all the data will be anonymized and will be used only for EASIER project.<br>**How it will be accessible:** EASIER repository and extracts of data in D1.1<br>**Methods/software tools to access data:** N/A<br>**Repository**: EASIER depository<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** No as long as if the extract of transcripts are shared. Trying to avoid sharing full transcripts outside of EASIER Consortium to maximise preservation of anonymization.<br>**Is informed consent for data sharing and long term preservation given?** Yes. Only for EASIER purpose and duration of EASIER project lifetime. |
| **Other issues** | N/A |

## 2.5 DATA GENERATED IN WP2

"WP2 does not make direct use of human signers' video resources. It receives coded input from WP3 which by no means reveals the identity of persons in videos. Similarly, WP2 work heavily depends on input from WP4 which involves the coded output of MT processes when translation direction is to SL. It needs to be mentioned that SL representation in WP2 does not face GDPR issues, since it does not involve any human data, but focuses on developing the various aspects of the project signing avatar."

## 2.6 DATA GENERATED IN WP3

*TABLE 3 EASIER_WP3_01_SRC_EMSL*

| **Identifier** | **EASIER_WP3_01_SRC_EMSL** |
|---|---|

| | |
|---|---|
| **Dataset Summary** | **Responsible partner:** UNIS<br>**Purpose**: Algorithms developed to extract spatio-temporal representations from sign language videos. These representations are essential for WP4's translation tasks.<br>**Type/format**: Source code, in python programming language. PEP8 is followed as a coding standard.<br>**Re-use of existing data**: We will be using the tools/datasets developed during the previous EU project, Content4All.<br>**Data origin**: Developed at University of Surrey.<br>**Expected size**: Source Code: Under 1GB. Models: Under 10GB.<br>**Data utility**: This source code will be useful for WP3 and WP4, for extracting representations from videos and using these representations to train translation models. |
| **Intellectual Property Rights** | **Owner:** University of Surrey<br>**Licensing:** No<br>**Dependency:** N/A<br>**Constraints:** It will be confidential to the University of Surrey.<br>**Timeframe:** It will be re-usable for the duration of the project and will be confidential to the University of Surrey. |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** No<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: They will be stored at the University of Surrey gitlab repositories. It will not be publicly accessible.<br>**Restrictions on access:** It will be confidential to the University of Surrey. |

Funded by the Horizon 2020
Framework Programme of the European Union

| | |
|---|---|
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** The source code will be stored at university of Surrey's secure servers. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** No<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | None |

## 2.7 DATA GENERATED IN WP4

*TABLE 4 EASIER_WP4_01_MUL_DSGS_BROADCAST*

| **Identifier** | **EASIER_WP4_01_MUL_DSGS_broadcast** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UZH<br>**Purpose**: These are pairings of Swiss German Sign Language videos and corresponding German subtitles from the Swiss Broadcasting Corporation. They are used to train the spoken-to-sign/sign-to-spoken translation models in the project. WP3 will generate European Meta Sign Language (EMSL) representations from the video side of the parallel corpus.<br>**Type/format**: The videos are available in MPEG-TS, the subtitles in SRT format.<br>**Re-use of existing data**: N/A<br>**Data origin**: Interpretations of the German daily news in Switzerland into Swiss German Sign Language<br>**Expected size**: approx. 4TB<br>**Data utility**: The data will primarily be useful to WP4. |
| **Intellectual Property Rights** | **Owner:** The data has been downloaded for research purposes by the relevant partners.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** Internally, metadata is specified in JSON format, storing information on file ID, TV channel, series, episode, date, time, description, duration, size, url, availability of subtitles, video width, video height, video framerate, video duration, biological sex of speakers, and biological sex of signers. |
| **Accessibility** | **Data openly accessible:** Copyright restrictions make publishing the data impossible. However, we are working on getting legal clearance to distribute a subset of the data at a later point in time.<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Data cannot be exchanged between partners in the project directly. However, the code for downloading the data is shared among partners. |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data.<br>**Is informed consent for data sharing and long term preservation given?** no |
| **Other issues** | N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

*TABLE 5 EASIER_WP4_02_MUL_LSF-CH_BROADCAST*

| Identifier | EASIER_WP4_02_MUL_LSF-CH_broadcast |
|---|---|
| **Dataset Summary** | **Responsible partner:** UZH<br>**Purpose**: These are pairings of French Sign Language of Switzerland videos and corresponding French subtitles from the Swiss Broadcasting Corporation. They are used to train the spoken-to-sign/sign-to-spoken translation models in the project. WP3 will generate European Meta Sign Language (EMSL) representations from the video side of the parallel corpus.<br>**Type/format**: The videos are available in MPEG-TS, the subtitles in SRT format.<br>**Re-use of existing data**: N/A<br>**Data origin**: Interpretations of the French daily news in Switzerland into French Sign Language of Switzerland<br>**Expected size**: approx. 3TB<br>**Data utility**: The data will primarily be useful to WP4. |
| **Intellectual Property Rights** | **Owner:** The data has been downloaded for research purposes by the relevant partners.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** Internally, metadata is specified in JSON format, storing information on file ID, TV channel, series, episode, date, time, description, duration, size, url, availability of subtitles, video width, video height, video framerate, video duration, biological sex of speakers, and biological sex of signers. |
| **Accessibility** | **Data openly accessible:** N/A |

| | |
|---|---|
| | **How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Data cannot be exchanged between partners in the project directly. However, the code for downloading the data is shared among partners. |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data.<br>**Is informed consent for data sharing and long-term preservation given?** no |
| **Other issues** | N/A |

*TABLE 6 EASIER_WP4_03_MUL_LIS-CH_BROADCAST*

| **Identifier** | **EASIER_WP4_03_MUL_LIS-CH_broadcast** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UZH<br>**Purpose**: These are pairings of Italian Sign Language of Switzerland videos and corresponding Italian subtitles from the Swiss Broadcasting Corporation. They are used to train the spoken-to-sign/sign-to-spoken translation models in the project. WP3 will generate European Meta Sign Language (EMSL) representations from the video side of the parallel corpus.<br>**Type/format**: The videos are available in MPEG-TS, the subtitles in SRT format.<br>**Re-use of existing data**: N/A<br>**Data origin**: Interpretations of the Italian daily news in Switzerland into Italian Sign Language of Switzerland<br>**Expected size**: approx. 2TB |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Data utility**: The data will primarily be useful to WP4. |
| **Intellectual Property Rights** | **Owner:** The data has been downloaded for research purposes by the relevant partners.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** Internally, metadata is specified in JSON format, storing information on file ID, TV channel, series, episode, date, time, description, duration, size, url, availability of subtitles, video width, video height, video framerate, video duration, biological sex of speakers, and biological sex of signers. |
| **Accessibility** | **Data openly accessible:** N/A<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Data cannot be exchanged between partners in the project directly. However, the code for downloading the data is shared among partners. |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Is informed consent for data sharing and long-term preservation given?** no |
| **Other issues** | N/A |

*TABLE 7 EASIER_WP4_04_MUL_DGS_BROADCAST*

| **Identifier** | **EASIER_WP4_04_MUL_DGS_broadcast** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: These are pairings of German Sign Language videos and corresponding German subtitles from German TV stations. They are used to train the spoken-to-sign/sign-to-spoken translation models in the project. WP3 will generate European Meta Sign Language (EMSL) representations from the video side of the parallel corpus.<br>**Type/format**: The videos are available in MPEG-4, the subtitles in SRT, VTT, or TTML format.<br>**Re-use of existing data**: N/A<br>**Data origin**: Interpretations of German shows into German Sign Language or original German Sign Language shows on German TV stations<br>**Expected size**: approx. 4TB<br>**Data utility**: The data will primarily be useful to WP4. |
| **Intellectual Property Rights** | **Owner:** The data has been downloaded for research purposes by the relevant partners.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Metadata creation:** Internally, metadata is specified in JSON format, storing information on file ID, TV channel, series, episode, date, time, description, duration, size, url, availability of subtitles, video width, video height, video framerate, video duration, biological sex of speakers, and biological sex of signers. |
| **Accessibility** | **Data openly accessible:** N/A<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Data cannot be exchanged between partners in the project directly. However, the code for downloading the data is shared among partners. |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data.<br>**Is informed consent for data sharing and long-term preservation given?** no |
| **Other issues** | N/A |

*TABLE 8 EASIER_WP4_05_MUL_BSL_BROADCAST*

| **Identifier** | **EASIER_WP4_05_MUL_BSL_broadcast** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UNIS<br>**Purpose**: These are pairings of British Sign Language videos and corresponding English subtitles from the BBC. They are used to train the spoken-to-sign/sign-to-spoken translation models in the project. WP3 will generate European Meta Sign Language (EMSL) representations from the video side of the parallel corpus.<br>**Type/format**: The videos are available in MPEG-4, the subtitles in SRT and TTML format. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Re-use of existing data**: N/A<br>**Data origin**: Interpretations of British shows into British Sign Language or original British Sign Language shows on the BBC<br>**Expected size**: approx. 4TB<br>**Data utility**: The data will primarily be useful to WP4. |
| **Intellectual Property Rights** | **Owner:** The data will be downloaded for research purposes by the relevant partners.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** Information on the biological sex of speakers and signers will be added as part of T6.3. |
| **Accessibility** | **Data openly accessible:** N/A<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Data cannot be exchanged between partners in the project directly. However, the code for downloading the data is shared among partners. |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data.<br>**Is informed consent for data sharing and long term preservation given?** no |
| **Other issues** | N/A |

*TABLE 9 EASIER_WP4_06_MUL_TRAINING_DATA_QUALITY_ESTIMATION_MODELS*

| **Identifier** | **EASIER_WP4_06_MUL_training_data_quality_estimation_models** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UZH<br>**Purpose**: This is the data used to train sentence-based automatic machine translation quality estimation (QE) models for German and German Sign Language.<br>**Type/format**: The data is represented as plain text files.<br>**Re-use of existing data**: N/A<br>**Data origin**: The data will consist of German sentences and their machine translations into German Sign Language, represented as EMSL (as generated in WP3), and vice versa.<br>**Expected size**: The data will comprise 5,000 examples and will be in the MB range.<br>**Data utility**: The data will be used by the partners in WPs 4 and 5. |
| **Intellectual Property Rights** | **Owner:** The data will be curated by UZH.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| **Accessibility** | **Data openly accessible:** The source sentences involved in the dataset are from a source that has not been legally cleared for release.<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data.<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | N/A |

## 2.8 DATA GENERATED IN WP5

*TABLE 10 EASIER_WP5_01_OTH_POSTEDITEDDATA*

| Identifier | EASIER_WP5_01_OTH_postEditedData |
|---|---|
| **Dataset Summary** | **Responsible partner:** STXT<br>**Purpose**: Train models in T4.5 from human post-edited translations.<br>**Type/format**: N/A<br>**Signing videos**: mp4, Container format provided by MPEG for MPEG-4 content and standardised in ISO/IEC 14496-12 and -14<br>**Text**: SRT, Segmented timecoded Text file format.<br>**Re-use of existing data**: No.<br>**Data origin**: Human input collected in the WP tasks. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Expected size**: 5,000 sentences + 5,000 signed entries<br>**Data utility**: WP4 and constitutes D5.5 |
| **Intellectual Property Rights** | **Owner:**<br>○ **SILAS: rights owed by STXT and the interpreters**<br>○ **Edit:  performed by WP5 participants.**<br>**Licensing:** tbd<br>**Dependency:** original Data of signing and editing (video & text ) owned by individual interpreters<br>**Constraints:** copyright issues between EU and CH<br>**Timeframe:**  tbd |
| **Findability** | **DOI:** No (confidential)<br>**Is data discoverable:** No (confidential)<br>**Search keywords:** No (confidential)<br>**Versioning:** Two versions, respectively constituting D5.5 and D5.6.<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** No (confidential)<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | N/A |

TABLE 11 EASIER_WP5_02_MUL_VERBALISINGDIAGRAMCORPUS

| Identifier | EASIER_WP5_02_MUL_verbalisingDiagramCorpus |
|---|---|
| **Dataset Summary** | **Responsible partner:** CNRS<br>**Purpose**: Study spontaneous productions of signers representing their language in a graphical form in T5.3<br>**Type/format**: Images (PNG or multi-page PDF) + video (MPEG4)<br>**Re-use of existing data**: This is an existing data set, with an option to extend it slightly in WP5<br>**Data origin**: Previously elicited corpus, as per Michael Filhol, “Elicitation and corpus of spontaneous Sign Language discourse representation diagrams”, in *Proceedings of the 9th workshop on the Representation and Processing of Sign Languages*, May 2020.<br>**Expected size**: ~500 GB<br>**Data utility**: T5.3, to specify the format edited in the implemented editor |
| **Intellectual Property Rights** | **Owner:** CNRS<br>**Licensing:** N/A<br>**Dependency:** N/A<br>**Constraints:** Data still undisclosed due to option to continue extending it.<br>**Timeframe:** When data complete and after a couple of years of exclusive use, data will be deposited on a research corpus data platform (except for the videos of informants who chose against it). |
| **Findability** | **DOI:** Yes.<br>**Is data discoverable:** Yes.<br>**Search keywords:** Verbalising diagrams, graphical SL representation.<br>**Versioning:** No.<br>**Metadata creation:** No standard. A simple field-value list will be created to describe the data. |
| **Accessibility** | **Data openly accessible:** Yes.<br>**How it will be accessible:** Research data repository (e.g. OrtoLang) |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Methods/software tools to access data:** PDF/image viewer and video player.<br>**Repository**: OrtoLang (tbc)<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Yes (standard image/video formats) |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** The video content shows people's faces, while some informants might choose not to be displayed online.<br>**Is informed consent for data sharing and long-term preservation given?** Yes. |
| **Other issues** | |

*TABLE 12 EASIER_WP5_03_MUL_QUALITYESTIMATIONTRAININGDATA*

| **Identifier** | **EASIER_WP5_03_MUL_qualityEstimationTrainingData** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UZH<br>**Purpose**: This is the data used to train sentence-based automatic machine translation quality estimation (QE) models for German and German Sign Language.<br>**Type/format**: The data is represented as plain text files.<br>**Re-use of existing data**: N/A<br>**Data origin**: The data will consist of German sentences and their machine translations into German Sign Language, represented as EMSL (as generated in WP3), and vice versa.<br>**Expected size**: The data will comprise 5,000 examples and will be in the MB range.<br>**Data utility**: The data will be used by the partners in WPs 4 and 5. |
| **Intellectual Property Rights** | **Owner:** The data will be curated by UZH.<br>**Licensing:** The data cannot be distributed.<br>**Dependency:** N/A |

| | |
|---|---|
| | **Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** The source sentences involved in the dataset are from a source that has not been legally cleared for release.<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** The data is stored on secured file servers. Access to the data within each partner institution is granted to members of the project team only and is managed through an identity management system. Any person accessing the data is a registered and verified user in the identity management system. Members of the project team will access data on the servers via secured networks. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** We do not have legal clearance to share the data.<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | N/A |

## 2.9 DATA GENERATED IN WP6

WP06 focuses on the harmonization of existing resources, i.e. there is no plan to do (field) data collections within this work package. Instead, WP6 will aggregate existing data in novel ways and, where necessary, close gaps. Filling gaps on linguistic detail will mostly be achieved by the language intuition of project team members of external consultants. For illustration purposes, this new data may be recorded as video, unavoidably resulting in bits of data that show human subjects. In these cases, we will make sure that co-workers are well aware of

Funded by the Horizon 2020 Framework Programme of the European Union

their rights so that they can either enter a specific contract with their employer or sign informed consents.

*TABLE 13 EASIER_WP6_01_MUL_LREC-ANTHOLOGY*

| **Identifier** | **EASIER_WP6_01_MUL_LREC-Anthology** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: The sign-lang@LREC anthology contains all papers published at the LREC workshop on sign languages which is the event where most papers on linguistic resources for sign languages are presented. It also includes select papers from the LREC main conference that address sign languages. The resource indexes the papers not only by conferences and authors, but also by the datasets and tools introduced/used, sign languages covered, and research projects that the paper originate from.<br>**Type/format**: Website<br>**Re-use of existing data**: This resource enriches a collection of existing papers with task-specific metadata that are also of general interest for the sign language data community.<br>**Data origin**: LREC workshop papers<br>**Expected size**: 360 papers, 700 MB<br>**Data utility**: Within EASIER, the data are used as the basis for the overviews of linguistic corpora and lexical resources on European sign languages (EASIER-06-3 and EASIER-06-4). Beyond the project, the resource can be expected to become the major point of orientation for any researcher working with sign language resources. |
| **Intellectual Property Rights** | **Owner:** The original papers are copyrighted by ELRA and the authors, the additionally created metadata will be in the public domain.<br>**Licensing:** The papers are licensed under a CC BY-NC license. The site itself will be free to use for everyone and licensed under an open license.<br>**Dependency:** As the website is non-commercial, re-use of papers is permitted.<br>**Constraints:** None<br>**Timeframe:** N/A |
| **Findability** | **DOI:** The resource as a whole will have a DOI. It is currently under discussion if the project will also provide DOIs for individual papers in the anthology. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Is data discoverable:** The website uses established metadata standards to ensure indexing on both general-purpose search engines as well as academic search engines (e.g. Google Scholar).<br>**Search keywords:** The resource as a whole is advertised with appropriate keywording, the individual entries show keywords assigned by the project as part of this data collection.<br>**Versioning:** The resource will be updated for corrections and when new events in the workshop series occur.<br>**Metadata creation:** The content of the resource as bibliographic data is available in bibtex format and via semantic web metadata standards (Highwire Press, Dublin Core, Eprints, Open Graph). |
| **Accessibility** | **Data openly accessible:** Data produced in the project will be made openly available.<br>**How it will be accessible:** The data is made available as a website.<br>**Methods/software tools to access data:** N/A<br>**Repository**: Parts of the data reside in a public research data repository, the website providing application-specific access mechanism resides outside the repository.<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Data are available in bibtex and semantic web formats (see above). |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | none |

*TABLE 14 EASIER_WP6_02 MUL_DATA_COLLECTION_TASKS*

| **Identifier** | **EASIER_WP6_02 MUL_Data_Collection_Tasks** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: This resource lists data collection tasks that have been used in known sign language corpus projects to serve as a basis for cross-referencing resources in order to identify shared tasks. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Type/format**: Website<br>**Re-use of existing data**: This resource identifies data collection tasks that have been used in more than one sign language corpus to allow the user to construct meta-collection with specific types of data.<br>**Data origin**: This resource builds on excerpts from dataset EASIER_06_1 as well as personal communication with data authors, so it is primarily a bibliographic, i.e. metadata resource.<br>**Expected size**: <10 MB<br>**Data utility**: Within EASIER, the data are used as the basis for the overviews of linguistic corpora on European sign languages (EASIER-06-3). Beyond the project, the resource can be expected to become a major point of orientation for any researcher interested in cross-linguistic sign language data. |
| **Intellectual Property Rights** | **Owner:** EASIER<br>**Licensing:** The site will be free to use for everyone<br>**Dependency:** As described, the resource refers to data collections and scientific papers<br>**Constraints:** None.<br>**Timeframe:** N/A |
| **Findability** | **DOI:** The resource as a whole as well as each individual description will have a DOI.<br>**Is data discoverable:** The website is public. Semantic web standards will be used to optimize search engine indexing.<br>**Search keywords:** The resource as a whole is advertised with appropriate keywording.<br>**Versioning:** The resource will be updated during the lifetime of the project, at the latest with the next sign-lang@LREC workshop.<br>**Metadata creation:** The resource consists of metadata only. |
| **Accessibility** | **Data openly accessible:** Data produced in the project will be made openly available.<br>**How it will be accessible:** The data is made available as a website.<br>**Methods/software tools to access data:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Repository**: Parts of the data reside in a public research data repository, the website providing application-specific access mechanism resides outside the repository.<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Existing metadata standards will be used where feasible. |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | none |

*TABLE 15 EASIER_WP6_03_MUL_CORPORA*

| **Identifier** | **EASIER_WP6_03_MUL_Corpora** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: This resource lists linguistic sign language corpora of European sign languages.<br>**Type/format**: tbd<br>**Re-use of existing data**: Metadata referring to the original data when useful.<br>**Data origin**: This resource builds on excerpts from dataset EASIER_06_1 as well as personal communication with data authors, so it is primarily a bibliographic, i.e. metadata resource.<br>**Expected size**: <10 MB<br>**Data utility**: Within EASIER, this resource lists the scope of data harmonization candidates together with the most important features of these corpora from the perspective of making these data high-quality training data for the EASIER machine translation pipeline. Beyond the project, the resource is most useful for researchers looking for existing sign language corpora. |
| **Intellectual Property Rights** | **Owner:** EASIER<br>**Licensing:** The resource will be free to use for everyone |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Dependency:** As described, the resource provides specific metadata for linguistic sign language corpora and refers to these corpora to view the data. It builds on data from dataset EASIER_06_1_MUL_LREC-Anthology and uses dataset EASIER_06_2_MUL_Data_Collection_Tasks<br>**Constraints:** None.<br>**Timeframe:** N/A |
| **Findability** | **DOI:** The resource as a whole will have a DOI.<br>**Is data discoverable:** yes, the approach depends on the exact format still to be defined.<br>**Search keywords:** The resource as a whole is advertised with appropriate keywording.<br>**Versioning:** The resource will be updated during the lifetime of the project, at the latest with the next sign-lang@LREC workshop.<br>**Metadata creation:** The resource consists of metadata only. |
| **Accessibility** | **Data openly accessible:** Data produced in the project will be made openly available.<br>**How it will be accessible:** The data may be made available as a website or as a document available on the project website.<br>**Methods/software tools to access data:** N/A<br>**Repository**: Depending on the final decision on the format, the resource itself or the underlying data will be submitted to a public research data repository.<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | none |

Funded by the Horizon 2020 Framework Programme of the European Union

*TABLE 16 EASIER_WP6_04_MUL_LEXICAL_RESOURCES*

| Identifier | EASIER_WP6_04_MUL_Lexical_Resources |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: This resource lists machine-readable lexical resources for European sign languages.<br>**Type/format**: tbd<br>**Re-use of existing data**: Metadata referring to the original data when useful.<br>**Data origin**: This resource builds on excerpts from dataset EASIER_06_1 as well as personal communication with data authors, so it is primarily a bibliographic, i.e. metadata resource.<br>**Expected size**: <10 MB<br>**Data utility**: Within EASIER, this resource lists the scope of candidate data for inclusion in the multi-lingual lexical resource to be built in EASIER. |
| **Intellectual Property Rights** | **Owner:** EASIER<br>**Licensing:** The resource will be free to use for everyone<br>**Dependency:** As described, the resource provides specific metadata for sign language lexical resources and refers to these resources to view the data.<br>**Constraints:** None.<br>**Timeframe:** N/A |
| **Findability** | **DOI:** The resource as a whole will have a DOI.<br>**Is data discoverable:** yes, the approach depends on the exact format still to be defined.<br>**Search keywords:** The resource as a whole is advertised with appropriate keywording.<br>**Versioning:** The resource will be updated during the lifetime of the project, at the latest with the next sign-lang@LREC workshop.<br>**Metadata creation:** The resource consists of metadata only. |

| | |
|---|---|
| **Accessibility** | **Data openly accessible:** Data produced in the project will be made openly available.<br>**How it will be accessible:** The data may be made available as a website or as a document available on the project website.<br>**Methods/software tools to access data:** N/A<br>**Repository**: Depending on the final decision on the format, the resource itself or the underlying data will be submitted to a public research data repository.<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**:  tbd |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | none |

*TABLE 17 EASIER_WP6_05_MUL_INTERLINGUAL_INDEX*

| **Identifier** | **EASIER_WP6_05_MUL_Interlingual_Index** |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: This interlingual index binds available lexical descriptions for European sign languages to an existing multi-lingual wordnet.<br>**Type/format**: tbd<br>**Re-use of existing data**: Builds on existing multi-lingual wordnet data and will link to lexical resources. The exact datasets involved are tbd.<br>**Data origin**: tbd<br>**Expected size**: <1 MB<br>**Data utility**: Within EASIER, this resource will be used in the multi-lingual automatic translation pipeline. Beyond the project, this resource will be of interest for computational sign linguists. |
| **Intellectual Property Rights** | **Owner:** EASIER |

| | |
|---|---|
| | **Licensing:** The resource will be free to use for everyone.<br>**Dependency:** The resource provides specific metadata for sign language lexical resources and refers to these resources to view the data. **The resource will use data from Open Multilingual Wordnet, which is under open licence.**<br>**Constraints:** None.<br>**Timeframe:** N/A |
| **Findability** | **DOI:** The resource as a whole will have a DOI.<br>**Is data discoverable:** yes, the approach depends on the exact format still to be defined.<br>**Search keywords:** The resource as a whole is advertised with appropriate keywording.<br>**Versioning:** The resource will be updated regularly during the lifetime of the project.<br>**Metadata creation:** tbd. |
| **Accessibility** | **Data openly accessible:** Data produced in the project will be made openly available.<br>**How it will be accessible:** The data will be available from the research data repository, cross-linked from the EASIER project webpages.<br>**Methods/software tools to access data:** tbd<br>**Repository**: The data will be deposited in a public research data repository.<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: tbd |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | none |

*TABLE 18 EASIER_WP6_06 SRC_DATA_HARMONIZATION_TOOLS*

| Identifier | EASIER_WP6_06 SRC_Data_Harmonization_Tools |
|---|---|
| **Dataset Summary** | **Responsible partner:** UHH<br>**Purpose**: Set of tools to allow the integration of language data (corpora, lexical resources) into the EASIER translation pipeline and EASIER_06_5_MUL_Interlingual_Index.<br>**Type/format**: tbd<br>**Re-use of existing data**: tbd<br>**Data origin**: N/A<br>**Expected size**: N/A<br>**Data utility**: EASIER will use these tools to integrate the language resources chosen from EASIER_06_3_MUL_Corpora and EASIER_06_4_MUL_Lexical_Resources. The tools will also be used by external parties to prepare the future integration of additional languages by external parties. |
| **Intellectual Property Rights** | **Owner:** EASIER.<br>**Licensing:** Open source license<br>**Dependency:** tbd<br>**Constraints:** No.<br>**Timeframe:** N/A |
| **Findability** | **DOI:** DOIs will be created as part of code archiving.<br>**Is data discoverable:** Code will be hosted in a repository for open source code and archived in a research data repository.<br>**Search keywords:** Repositories will be given appropriate keywords.<br>**Versioning:** Semantic versioning will be used.<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** The tools will be open source.<br>**How it will be accessible:** Open source code repository.<br>**Methods/software tools to access data:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Repository**: tbd<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: Code will be open-source. Requirements for data input and data or code dependencies (if any) are tbd. |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | none |

## 2.10 DATA GENERATED IN WP7

*TABLE 19 EASIER_WP7_01_MUL_ANNOTATED_VIDEO_DATA*

| **Identifier** | **EASIER_WP7_01_MUL_annotated_video_data** |
|---|---|
| **Dataset Summary** | **Responsible partner:** [DFKI]<br>**Purpose**: These are the datasets used to train the vision-based affect recognition models for German and German Sign Language.<br>**Type/format**: The data consists of video recordings (MPEG-4) with corresponding textual annotations.<br>**Re-use of existing data**: N/A<br>**Data origin**: The data is based on the Content4All dataset released as part of the corresponding EU project.<br>**Expected size**: 36GB<br>**Data utility**: The data will serve to train the machine-learning-based affect recognition models that are the target of WP7. |
| **Intellectual Property Rights** | **Owner:** N/A<br>**Licensing:** N/A<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** N/A<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** N/A |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** N/A<br>**Is informed consent for data sharing and long term preservation given?** N/A |
| **Other issues** | N/A |

## 2.11 DATA GENERATED IN WP8

*TABLE 20 EASIER_WP8_01_MUL_VIDEORECORDING*

| **Identifier** | **EASIER_WP8_01_MUL_VideoRecording** |
|---|---|
| **Dataset Summary** | **Responsible partner:** NURO<br>**Purpose**: Video recordings of Sign Input are used as source materials for the translation. The data is also used in speech input since the videos is used for affect recognition. The video files will be sent to the Video Recognition Engine and the Multimodal Affect Recognition Engine.<br>**Type/format**: MP4<br>**Re-use of existing data**: No<br>**Data origin**: User |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Expected size**: Depends on length and resolution quality of the used hardware but presumably in the range of 10 MB to 1 GB per video recording.<br>**Data utility**: All partners involved in working with the video recognition engine and the translation process will need this dataset. |
| **Intellectual Property Rights** | **Owner:** User<br>**Licensing:** No<br>**Dependency:** This is private data which will not be re-used per se.<br>**Constraints:** No<br>**Timeframe:** N/A |
| **Findability** | **DOI:** Yes (tbd)<br>**Is data discoverable:** Probably not since it will be only used for the translation process.<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** It is unknown at this point if any metadata is needed and needs to be attributed to each individual file. |
| **Accessibility** | **Data openly accessible:** No, this is private user data.<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** Data transfer will be secured through TSL/SSL encryption. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Since this is private data the consortium will not share this data besides the intended usage of translation.<br>**Is informed consent for data sharing and long-term preservation given?** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| Other issues | In general this data not be stored or processed, in fact it will be only forwarded to WP3 and WP7 |

*TABLE 21 EASIER_WP8_02_MUL_AUDIORECORDING*

| Identifier | EASIER_WP8_02_MUL_AudioRecording |
|---|---|
| Dataset Summary | **Responsible partner:** NURO<br>**Purpose**: Audio recordings of Speech Input are used as source materials for the translation. The video files will be sent to the Speech Recognition Engine and the Multimodal Affect Recognition Engine.<br>**Type/format**: MP4<br>**Re-use of existing data**: No<br>**Data origin**: User<br>**Expected size**: Depends on length and resolution quality of the used hardware but presumably less than 10 MB per audio recording.<br>**Data utility**: All partners involved in the speech recognition engine will need this data set. |
| Intellectual Property Rights | **Owner:** User<br>**Licensing:** No<br>**Dependency:** This is private data which will not be re-used per se.<br>**Constraints:** No<br>**Timeframe:** N/A |
| Findability | **DOI: No**<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| Accessibility | **Data openly accessible:** No, this is private user data.<br>**How it will be accessible:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** Data transfer will be secured through TSL/SSL encryption. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Since this is private data the consortium will not share this data besides the intended usage of translation.<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | In general this data not be stored or processed, in fact it will be only forwarded to WP3 and WP7 |

*TABLE 22 EASIER_WP8_03_PER_USERDATASET*

| **Identifier** | **EASIER_WP8_03_PER_UserDataset** |
|---|---|
| **Dataset Summary** | **Responsible partner:** [NURO]<br>**Purpose**: A set of personal data is not important for the first phase of development. Later, however, against the background of an intended fully developed application, a set of personal data will be used for registration. In any case, a valid email address is needed to confirm the registration in the app and the login. The extent to which further personal data may be required is still to be clarified.<br>**Type/format**: Text<br>**Re-use of existing data**: No<br>**Data origin**: User<br>**Expected size**: MB range<br>**Data utility**: Unknown if any other partners can benefit from this data. |
| **Intellectual Property Rights** | **Owner:** User<br>**Licensing:** N/A<br>**Dependency:** This is private data which will not be re-used per se. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** Probably not since it will be only used for the translation process.<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** It is unknown at this point if any metadata is needed and needs to be attributed to each individual file. |
| **Accessibility** | **Data openly accessible:** No, this is private user data.<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** N/A<br>**Repository**: N/A<br>**Restrictions on access:** N/A |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** Data transfer will be secured through TSL/SSL encryption. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Since this is private data the consortium will not share this data besides the intended usage of translation.<br>**Is informed consent for data sharing and long-term preservation given?** N/A |
| **Other issues** | In general this data not be stored or processed, in fact it will be only forwarded to WP3 and WP7 |

## 2.12 DATA GENERATED IN WP9

*TABLE 23 EASIER_WP9_01_PER_CONTACTLIST*

| **Identifier** | **EASIER_WP9_01_PER_ContactList** |
|---|---|
| **Dataset Summary** | **Responsible partner:** RU |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Purpose**: A list of contacts will be established at European institutes that are currently in the process of creating corpora and/or lexical datasets for sign languages that are not part of the EASIER set of sign languages.<br>**Type/format**: CSV table<br>**Re-use of existing data**: No<br>**Data origin**: Collected online and through personal networks<br>**Expected size**: Few KBs<br>**Data utility**: It will help strengthen the network of sign language (technology) researchers in Europe. |
| **Intellectual Property Rights** | **Owner:** Public domain<br>**Licensing:** Fully open access<br>**Dependency:** No<br>**Constraints:** No<br>**Timeframe:** No embargo after project end |
| **Findability** | **DOI:** No<br>**Is data discoverable:** No<br>**Search keywords:** No<br>**Versioning:** One release at end of project<br>**Metadata creation:** None |
| **Accessibility** | **Data openly accessible:** Yes<br>**How it will be accessible:** EASIER website; RU website<br>**Methods/software tools to access data:** No special software needed<br>**Repository**: EASIER website, RU website<br>**Restrictions on access:** N/a |
| **Interoperability** | **Interoperability**: Yes |
| **Security** | **Security measures:** No |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** No |

| | |
|---|---|
| | **Is informed consent for data sharing and long term preservation given?** Informed consent will be obtained from all contacts appearing in the list. |
| **Other issues** | N/A |

## 2.13 DATA GENERATED IN WP10

*TABLE 24 EASIER_WP10_01_PER_NEWSLETTER*

| **Identifier** | **EASIER_WP10_01_PER_NEWSLETTER** |
|---|---|
| **Dataset Summary** | **Responsible partner:** MARTEL<br>**Purpose**: The personal data collected as part of a registration for the EASIER newsletter will only be used to verify registrations and to send our newsletter.<br>**Type/format:** E-mail address of visitors registering to the newsletter. During the registration for the newsletter, we also store the IP address of the computer system assigned by the Internet service provider (ISP) and used by the data subject at the time of the registration, as well as the date and time of the registration. The newsletter of EASIER contains so-called tracking pixels. A tracking pixel is a miniature graphic embedded in such e-mails, which are sent in HTML format to enable log file recording and analysis. Based on the embedded tracking pixel, Martel may see if and when an e-mail was opened by a data subject, and which links in the e-mail were called up by data subjects.<br>**Re-use of existing data**: No existing data is re-used.<br>**Data origin**: The personal data is collected previous informed consent upon registration to the newsletter through the dedicated page of EASIER's website.<br>**Expected size**: Between 100 and 200 records.<br>**Data utility**: Data collected express the interest of subscribers to receive news from EASIER project. |
| **Intellectual Property Rights** | **Owner:** In compliance with GDPR, data subjects are the only owner of their data and at any time entitled to revoke the respective separate declaration of consent issued by means of the double-opt-in procedure. After a revocation, these personal data will be deleted by Martel. Martel automatically regards a withdrawal from the receipt of the newsletter as a revocation. At the end of the project, data will be deleted.<br>**Licensing:** N/A<br>**Dependency:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** Data collected are personal data (email address and other contact details) as such they will not be shared with any third party (openly or privately).<br>**How it will be accessible:** N/A<br>**Methods/software tools to access data:** Access to the newsletter subscribers list occurs via username and password. Only authorised users have the access and only admins in Martel can grant access.<br>**Repository**: We use the third-party service provider Mailchimp (The Rocket Science Group, LLC), to process and store your data: which EASIER uses to manage the newsletter subscriber lists and send emails to our subscribers. We don't share users data with any other third-party.<br>EASIER Privacy Policy is linked from each newsletter Users data are not used after the end of the project for marketing purposes.<br>**Restrictions on access:** Only people authorised by the data owner (MARTEL) will have access to the collected data. The data owner can authorise as data processors other project partners involved with the dissemination efforts of the project only where necessary. |
| **Interoperability** | **Interoperability**: N/A |
| **Security** | **Security measures:** Data are stored in a GDPR compliant newsletter service. Data are encrypted, and access to them is possible only via authentication of previously authorised users by the data owner entity (MARTEL). Authentication and access channels to the data are as well encrypted. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Data collected are personal data, and in compliance with GDPR can only be used for the original scope they have been collected for: informing newsletter subscribers about EASIER. As such they will not be shared outside the parties in the consortia in charge of providing the newsletter services (so not to any external third party). |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Is informed consent for data sharing and long-term preservation given?** All related information is available in detail at the following URL (on section 5 and 6): https://www.project-easier.eu/privacy-policy.<br>The subscription to our newsletter may be terminated by the data subject at any time. The consent to the storage of personal data, which the data subject has given for shipping the newsletter, may be revoked at any time. For the purpose of revocation of consent, a corresponding link is found in each newsletter. It is also possible to unsubscribe from the newsletter at any time by contacting Martel/EASIER's consortium directly via the Contact section of the website, or through the website's GDPR Requests page: https://www.project-easier.eu/privacy-policy.<br>Furthermore, subscribers to the newsletter may be informed by e-mail, as long as this is necessary for the operation of the newsletter service or a registration in question, as this could be the case in the event of modifications to the newsletter offer, or in the event of a change in technical circumstances. |
| **Other issues** | N/A |

*TABLE 25 EASIER_WP10_02_PER_WEBSITE*

| **Identifier** | **EASIER_WP10_02_PER_WEBSITE** |
|---|---|
| **Dataset Summary** | **Responsible partner:** MARTEL<br>**Purpose**: The data collected from the website visitors is needed for maintaining contact, and as required by law for any legal agreement we have with them. The data and information collected are used to optimize content and performance of the website and may also be used in the event of attacks on our information technology systems.<br>**Type/format**: The web site collects the following data:<br>• User account data: i.e. account of users authorised to publish content on the web site. This are in general users part of MARTEL and in some cases of other third parties in the EASIER authorised to publish news on the web site. This may include:<br>○ Name and relevant titles<br>○ Email<br>○ Job title<br>○ Company name<br>○ Company address (in very rare cases)<br>• Contact data: i.e. data provided by web sites visitors filling in contact forms to request information to the EASIER consortia. This may include:<br>○ Name and relevant titles<br>○ Email<br>○ Job title<br>○ Company name |

| | |
|---|---|
| | ○ Company address (in very rare cases)<br>• Access log data: i.e. data that are collected by servers to control access to the web pages of the web site that may be used in the event of attacks and other cases requested by the law. This may include (1) the browser types and versions used, (2) the operating system used by the accessing system, (3) the website from which an accessing system reaches our website (so-called referrers), (4) the sub-websites, (5) the date and time of access to the Internet site, (6) an Internet protocol address (IP address), (7) the Internet service provider of the accessing system, and (8) any other similar data and information that may be used in the event of attacks on our information technology systems.<br>**Re-use of existing data**: No existing data is re used.<br>**Data origin**: All data we store has been clearly and voluntarily provided to us by contacts and partners (subscribing to a newsletter, sending us an email, giving us their business card, answering to an online questionnaire etc.).<br>• The website collects a series of general data and information when a data subject or automated system calls up the website.<br>• The website contains information that enables a quick electronic contact to the project's consortium, as well as direct communication with EASIER, which also includes a general address of the so-called electronic mail (e-mail address). If a data subject contacts the project by e-mail or via a contact form, the personal data transmitted by the data subject are automatically stored. Such personal data transmitted on a voluntary basis by a data subject to Martel as data owner are stored for the purpose of processing or contacting the data subject.<br>As is common practice with almost all professional websites this site also uses cookies, to improve visitors' experience. In some special cases we also use cookies anonymized provided by trusted third parties (such as Google Analytics) – More details are provided in the Cookie Policy page of EASIER's website: [https://www.project-easier.eu/cookie-policy](https://www.project-easier.eu/cookie-policy).<br>**Expected size**: We collect minimal data from our site visitors.<br>**Data utility**: Data and information collected are needed to (1) deliver the content of our website correctly, (2) optimize the content of our website as well as its advertisement, (3) ensure the long-term viability of our information technology systems and website technology, and (4) provide law enforcement authorities with the information necessary for criminal prosecution in case of a cyber-attack. |
| **Intellectual Property Rights** | **Owner:** N/A<br>**Licensing:** N/A<br>**Dependency:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A<br>**Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** Data collected are personal data (email address and other contact details) as such they will not be shared with any third party (openly or privately).<br>**How it will be accessible:** The data owner (MARTEL) can authorise as data processors other project partners involved with the dissemination efforts of the project only where necessary.<br>**Methods/software tools to access data:** Access to the web site log data occurs via username and password. Only authorised users have the access and only admins in Martel can grant access.<br>**Repository**: See Security section.<br>**Restrictions on access:** Only people authorised by the data owner (MARTEL) will have access to the collected data. |
| **Interoperability** | **Interoperability**: There will be no transfer of personal data collected to any third party. |
| **Security** | Users can access the website only via an encrypted connection (https), in order to add a second security layer between the user and the website itself.<br>WordPress (WP) has been used to build EASIER's website. This content management system (CMS) uses the latest technology about PHP and MariaDB for the business logic and database respectively. WP provides al lot of plug ins in order to grant a great security both for the content and users. In fact, plug ins such as anti-spam, anti-SQL injection, anti-brute force attack etc. can help to prevent spam and the most common attacks. Moreover, WP provides different access roles, in order to grant the right permissions to the right users. Users' WP passwords are encrypted through RSA technology, so no one can decrypt them. Neither a WP administrator.<br>Setting of cookies can be prevented by visitor by adjusting the settings on their browser. Disabling cookies will usually result in also disabling |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | certain functionality and features of this site. Therefore, it is recommended for users not disable cookies.<br>On this website, Martel has integrated components of Twitter: Twitter receives information via the Twitter component that the data subject has visited our website, provided that the data subject is logged in on Twitter at the time of the call-up to our website. This occurs regardless of whether the person clicks on the Twitter component or not. If such a transmission of information to Twitter is not desirable for the data subject, then he or she may prevent this by logging off from their Twitter account before a call-up to our website is made. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Data collected are personal data, and in compliance with GDPR can only be used for the original scope they have been collected for: informing contacts about EASIER. As such they will not be shared outside the parties in the consortia in charge of providing the newsletter services (so not to any external third party).<br>**Is informed consent for data sharing and long term preservation given?** The website respects the latest European laws about Privacy (GDPR). All users (registered and guests) have to authorise data collection by the data owner (MARTEL) and eventually, in the case of cookies for anonymised tracking – the sharing with a third party. All information is available in detail at the following URLs:<br>- https://www.project-easier.eu/privacy-policy<br>- https://www.project-easier.eu/cookie-policy<br>The aforementioned Privacy Policy page also informs visitors in detail on the data subject's rights (Section 9). |
| **Other issues** | Martel shall process and store the personal data of the data subject only for the period necessary to achieve the purpose of storage, or as far as this is granted by the European legislator or other legislators in laws or regulations to which the Martel is subject to.<br>If the storage purpose is not applicable, or if a storage period prescribed by the European legislator or another competent legislator expires, the personal data are routinely blocked or erased in accordance with legal requirements.<br>The criteria used to determine the period of storage of personal data is the respective statutory retention period. After expiration of that period, the corresponding data is routinely deleted. |

Funded by the Horizon 2020 Framework Programme of the European Union

## 2.14 DATA GENERATED IN WP11

*TABLE 26 EASIER_WP11_01_PER_MATTERMOST*

| Identifier | EASIER_WP11_01_PER_Mattermost |
|---|---|
| **Dataset Summary** | **Responsible partner:** MARTEL<br>**Purpose**: The data collected within the Mattermost platform (open source tool installed on a private cloud managed by Martel) are solely used for enabling the user the access to the platform and allowing the usage of the messaging service. Data generated within the platform are private to the consortium.<br>**Type/format**: The Mattermost online application collects the following personal data:<br>• Username<br>• Nickname (optional)<br>• Position (optional)<br>• Email<br>• Profile picture (optional)<br>**Re-use of existing data**: No existing data is re used.<br>**Data origin**: All data we store has been clearly and voluntarily provided to us by consortium partners. Mattermost is an online application which present itself as a website allowing group and direct communication withing EASIER.<br>**Expected size**: We collect minimal data from our site visitors.<br>**Data utility**: Data and information collected are needed to allow the user to use the online tool. |
| **Intellectual Property Rights** | **Owner:** N/A<br>**Licensing:** N/A<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** Data collected are personal data of consortium members (email address and other contact details) as such they will not be shared with any third party (openly or privately).<br>**How it will be accessible:** The data owner (MARTEL) will not disclose any of the provided data.<br>**Methods/software tools to access data:** Only authorised users in Martel have the access and only admins in Martel can grant access.<br>**Repository**: See Security section.<br>**Restrictions on access:** Only authorised users in Martel have the access and only admins in Martel can grant access. |
| **Interoperability** | **Interoperability**: There will be no transfer of personal data collected to any third party. |
| **Security** | Users can access the Mattermost service only via an encrypted connection (https), in order to add a second security layer between the user and the website itself.<br>Mattermost has been used to build EASIER's dedicated messaging service. This messaging system uses the latest technology about GO, Javascript and PostgreSQL for the business logic and database respectively. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Data collected are personal data, and in compliance with GDPR can only be used for the original scope they have been collected for: internal consortium communication. As such they will not be shared outside the parties in the consortium.<br>**Is informed consent for data sharing and long-term preservation given?** Mattermost respects the latest European laws about Privacy (GDPR). Participants can cancel their subscription at any time. |
| **Other issues** | Martel shall process and store the personal data of the data subject only for the period necessary to achieve the purpose of storage, or as far as this is granted by the European legislator or other legislators in laws or regulations to which the Martel is subject to.<br>If the storage purpose is not applicable, or if a storage period prescribed by the European legislator or another competent legislator expires, the personal data are routinely blocked or erased in accordance with legal requirements. |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | The criteria used to determine the period of storage of personal data is the respective statutory retention period. After expiration of that period, the corresponding data is routinely deleted. |

TABLE 27 EASIER_WP11_02_PER_MAILMAN

| **Identifier** | **EASIER_WP11_02_PER_Mailman** |
|---|---|
| **Dataset Summary** | **Responsible partner:** MARTEL<br>**Purpose**: The email collected within the Mailman platform (open source mailing list tool installed on a private cloud managed by Martel) are solely used for enabling the user the access to the platform and allowing the usage of the messaging service. Data generated within the platform are private to the consortium.<br>**Type/format**: The Mailm online application collects the following personal data:<br>• Username (optional)<br>• Email<br>**Re-use of existing data**: No existing data is re used.<br>**Data origin**: All data we store has been clearly and voluntarily provided to us by consortium partners. Mailman is an online application which allow the management of the project’s mailing lists and has a light graphical interface for the users to subscribe or unsubscribe.<br>**Expected size**: We collect minimal data from our site visitors, about 50 entries overall, spread and/or duplicated within the different mailing lists (one per WP, general one,, etc)<br>**Data utility**: Data and information collected are needed to allow the user to use the online tool. |
| **Intellectual Property Rights** | **Owner:** N/A<br>**Licensing:** N/A<br>**Dependency:** N/A<br>**Constraints:** N/A<br>**Timeframe:** N/A |
| **Findability** | **DOI:** N/A<br>**Is data discoverable:** N/A<br>**Search keywords:** N/A |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Versioning:** N/A<br>**Metadata creation:** N/A |
| **Accessibility** | **Data openly accessible:** Data collected are personal data of consortium members (email address) as such they will not be shared with any third party (openly or privately).<br>**How it will be accessible:** The data owner (MARTEL) will not disclose any of the provided data.<br>**Methods/software tools to access data:** Only authorised users in Martel have the access and only admins in Martel can grant access.<br>**Repository**: See Security section.<br>**Restrictions on access:** Only authorised users in Martel have the access and only admins in Martel can grant access. |
| **Interoperability** | **Interoperability**: There will be no transfer of personal data collected to any third party. |
| **Security** | Users can access the Mailman admin interface only via an encrypted connection (https), in order to add a second security layer between the user and the website itself.<br>Mailman has been used to build EASIER's dedicated mailing lists service. This system uses the latest technology about Python. |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Data collected are personal data, and in compliance with GDPR can only be used for the original scope they have been collected for: internal consortium communication. As such they will not be shared outside the parties in the consortium.<br>**Is informed consent for data sharing and long term preservation given?** Mailman respects the latest European laws about Privacy (GDPR). Participants can cancel their subscription at any time. |
| **Other issues** | Martel shall process and store the personal data of the data subject only for the period necessary to achieve the purpose of storage, or as far as this is granted by the European legislator or other legislators in laws or regulations to which the Martel is subject to.<br>If the storage purpose is not applicable, or if a storage period prescribed by the European legislator or another competent legislator expires, the personal data are routinely blocked or erased in accordance with legal requirements.<br>The criteria used to determine the period of storage of personal data is the respective statutory retention period. After expiration of that period, the corresponding data is routinely deleted. |

Funded by the Horizon 2020 Framework Programme of the European Union

# 3 CONCLUSIONS

This deliverable provides an initial overview of the data that EASIER project will produce or envision to produce, together with the related data processes and requirements that need to be taken into consideration.

It is foreseen that most of the generated data will be two folds: source-code and datasets, which will follow standardised formats, and will be documented, so that they can be easily used as reference in other research projects or initiatives. The datasets will be disseminated through research publications, patent applications, invited talks, among others, and will be preserved at the partners repositories for at least three years. For completeness, also administrative datasets have been included in the DMP.

It should be noted that since it is very early in the project, this document only presents preliminary proposals in terms of sharing, volume and archiving, as well as DOI. The DMP will be updated periodically during the project (M18, M36) to reflect changes in the properties of the data that may be made available along the project, and to add more concrete information about the datasets.

Funded by the Horizon 2020
Framework Programme of the European Union

# APPENDIX A: LEGAL FRAMEWORK

The collection, use and disclosure of personal data at a European level are regulated in chronological order by the following:

- in 1995 by the 95/46/EC “Data Protection Directive”
- in 2002 by the 2002/58/EC “Privacy and electronic communications Directive”
- in 2009 by the 2009/136/EC “Cookie Directive”
- in 2016 by the 2016/679/EC “General Data Protection Regulation (GDPR)” that repeals Directive 95/46/EC
- in 2016 by the 2016/680/EC “Protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data Directive”.

Funded by the Horizon 2020 Framework Programme of the European Union

# APPENDIX B: TEMPLATE FOR DATA COLLECTION

Each WP Leader should collect the datasets intended to be generated in their WP, collecting information from the partners. In case a dataset might be shared among different WPs, the involved WP leaders should communicate offline and decide a collocation inside a single WP. The dependency should then be indicated in the table below: **Dataset Summary → Purpose.**

## IDENTIFIERS

Identify each data[set] with the format:

EASIER_<WP number>_<serial number of dataset>_<data type>_<dataset title>

Replace with your WP number as in the DoA, refer to the table below (Table 1) for the data type and assign a meaningful title.

*TABLE 28 IDENTIFIER PATTERN AND TYPES OF DATA*

| Acronym | Data Type |
|---|---|
| SRC | Source Code |
| MUL | Multimedia content (audio, video, text) |
| PER | Personal data (email addresses, other personal information) |
| OTH | Other type of data (to be specified in the table below) |

## DATASETS

Create a table using the provided template (Table 29) for each data[set] following the questions.

- ➲ Do you plan to generate/gather data (audio, video, text, logs, transcripts, parameters, algorithms, source code, software stacks, APIs, policies, etc.) inside the project or during the project lifetime in your WP?
  *<YES/NO, why>*
- ➲ If **yes**, please fill in. the table below answering the given questions (in case of not applicability, indicate N/A).

*TABLE 29 DATASET PROPERTIES TEMPLATE*

| Identifier | EASIER_<WP number>_<serial number of dataset>_<data type>_<dataset title> |
|---|---|
| **Dataset Summary** | **Responsible partner:** [MARTEL, ATHENA, UHH, RU, UNIS, UZH, CNRS, DFKI, AUD, NURO, STXT, EUD, INT, UCL] |

Funded by the Horizon 2020 Framework Programme of the European Union

| | |
|---|---|
| | **Purpose**: Short description of data. Also, what is the purpose of data collection/generation (and its relation to project objectives) in the context of EASIER and the Workpackage? Is this dataset being used in others EASIER WPs?<br>**Type/format**: What is the type/format of the data? Do you follow any standards for such data? (NIST, ISO, etc.)?<br>**Re-use of existing data**: Do you use any existing data/datasets, which you can/will reuse for further developments/implementations inside the project or during the project lifetime?<br>**Data origin**: What is the origin/source of the data?<br>**Expected size**: What is the expected data/dataset size? GB? TB?<br>**Data utility**: To whom will this data be useful and how? (inside the project and also to third parties, if applicable) |
| **Intellectual Property Rights** | **Owner:** Who owns the data?<br>**Licensing:** Will the data be licensed for reuse? If yes, under which licence?<br>**Dependency:** Are there any restrictions on the reuse of existing/third-party data?<br>**Constraints:** Will data sharing be postponed / restricted e.g. to publish or seek patents?<br>**Timeframe:** Specify the length of time for which the data will remain re-usable or if there is any embargo that prevent immediate publication. |
| **Findability** | **DOI:** Will you pursue getting a persistent identifier for your data?<br>**Is data discoverable:** Are the data produced in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?<br>**Search keywords:** Will search keywords be provided that optimize possibilities for re-use?<br>**Versioning:** Will clear version numbers be provided? How many releases are foreseen?<br>**Metadata creation:** Specify standards for metadata creation (if any). If there are no standards in your discipline, describe what type of metadata will be created and how. |
| **Accessibility** | **Data openly accessible:** Will data produced in the project be made openly available as the default? If certain datasets cannot be shared (or |

| | |
|---|---|
| | need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.<br>**How it will be accessible:** How will the data be made accessible (e.g. by deposition in an open repository)?<br>**Methods/software tools to access data:** What methods or software tools are needed to access the data? Also, is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?<br>**Repository**: Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.<br>**Restrictions on access:** If there are restrictions on use, how will access be provided? |
| **Interoperability** | **Interoperability**: Are the data produced in the project interoperable, i.e. can data be exchanged and re-used between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)? If not, what are the barriers and how to overcome them? |
| **Security** | **Security measures:** Security measures implemented for data protection (incl. controlled access, user authentication, firewalls, VPNs, encryption, back-ups, etc.) |
| **Ethical aspects** | **Possible ethical and legal aspects preventing sharing:** Are there any ethical or legal issues that can have an impact on data sharing? If yes, describe them and the way they are dealt with.<br>**Is informed consent for data sharing and long-term preservation given?** Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data? |
| **Other issues** | Refer to other national/funder/sectorial/departmental procedures for data management that you may be using (if any) |

Funded by the Horizon 2020 Framework Programme of the European Union