SIGNON

Sign Language Translation Mobile Application and Open Communications Framework

Deliverable 7.10: Data Management Report

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255

Project Information
Project Number: 101017255
Project Title: SignON: Sign Language Translation Mobile Application and Open Communications Framework
Funding Scheme: H2020 ICT-57-2020
Project Start Date: January 1st 2021

Deliverable Information
Title: Data Management Report
Work Package: WP 7 - Coordination and Management
Lead beneficiary: DCU
Due Date: 31/12/2023
Revision Number: V0.1
Authors: Aoife Brady, Henk van den Heuvel
Dissemination Level: Public
Deliverable Type: ORDP

Overview: This deliverable D7.10 provides an overview of the procedure that was applied to the data management process in the implementation of the SignON project. In the annexes it contains the final versions of the DMPs of the SignON partners.

Revision History

Version #	Implemented by	Revision Date	Description of changes
V0.1	Henk van den Heuvel	05/12/2023	First draft

The SignON project has received funding from the European Union's Horizon 2020 Programme under Grant Agreement No. 101017255. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the SignON project or the European Commission. The European Commission is not liable for any use that may be made of the information contained therein.

The Members of the SignON Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the SignON Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Approval Procedure

Version #	Deliverable Name	Approved by	Institution	Approval Date
V0.1	D7.10	Shaun O'Boyle	DCU	7.12.23
Vx.x	D7.10	Vincent Vandeghinste	INT	6.12.23
V0.1	D7.10	Adrián Núñez-Marcos	UPV/EHU	11.12.23
V0.1	D7.10	John O'Flaherty	MAC	06/12/2023
Vx.x	D7.10	Josep Blat	UPF	06/12/2023
V0.1	D7.10	Irene Murtagh	TU Dublin	14/12/23
Vx.x	D7.10	Caro Brosens	VGTC	06/12/2023
V0.1	D7.10	Henk van den Heuvel	RU	08/12/2023
Vx.x	D7.10	Lien Soetemans Myriam Vermeerbergen	KU Leuven	6/12/2023 10/12/2023
V0.1	D7.10	Davy Van Landuyt	EUD	08/12/2023
Vx.x	D7.10	Mirella De Sisto	TiU	7/12/2023

Acronyms

The following table provides definitions for acronyms and terms relevant to this document.

Acronym	Definition
APC	Article Processing Charge
DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulation
DHH	Deaf and Hard of hearing
FLOSS	Free/libre and Open-Source Software
RDM	Research Data Management
SL	Sign Language
Term	**Definition**
Data subjects	Legal term for participants included in data collections, in the case of SignON typically (but not only) from the DHH population.

Table of Contents

1. Executive Summary

Data Management Plans (DMPs) are a key element of good data management. In the definition of the EC,[1] a DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and reusable (FAIR), a DMP should include information on:

- the handling of research data during & after the end of the project,
- what data will be collected, processed and/or generated,
- which methodology & standards will be applied,
- whether data will be shared/made open access and,
- how data will be curated & preserved (including after the end of the project).

D7.8 (SignON Data Management Plan) contained the framework and requirements for the Data Management Plans of each project partner. In this deliverable, D7.10, we reflect on the process and provide the final DMP of each partner in the appendices.

2. Summary of Data Management in the SignON Consortium

A Data Management Team was established for the SignON Project and, with their help, the Data Management Plan (D7.8) was created and delivered in June 2021. That deliverable set out the principles for Research Data Management in the SignON project. It addressed the project's Open Access publication policy and how the project will meet the FAIR principles for sharing data, models and software after the lifetime of the project. It also explains how the collection and sharing of data (in a GDPR compliant way) during and after the lifetime of the project will be dealt with, as well as data security issues and ethical aspects.

A DMP is a dynamic document and it was decided that multiple DMPs will be needed by the various partners in the SignON project. The DMP's "living document" approach complements the GDPR

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

requirements of Privacy by Design and Privacy by Default. In other words, ensuring data protection is incorporated at the outset and throughout the project.

D7.8 was then complemented by the specific DMPs for each partner. These are included as an annex to this document and accommodated for the dynamic nature of DMPs. Intermediate versions of these DMPs were scheduled for M12, M24 & M36 and were reviewed by SignON's Research Ethics Committee (REC), as introduced in D9.1, to ensure compliance with the principles set out in D7.8.

Data Protection training was provided to all partners in May 2021 by the DCU Data Protection Unit. For those who could not attend, recordings were shared in the project Google drive and links are provided on the Intranet site.

In June 2021, a template for the Data Transfer Agreement (D8.5) was created and shared to lay out the conditions of the transfer of data between consortium members. In collaboration with the DCU Data Protection Unit, it was agreed that all partners would be data controllers and this agreement would be signed by all partners. Negotiations for the exact wording for this were lengthy but it was fully executed in 2023.

In D7.8 B2DROP and B2SHARE were offered as options for data and model sharing during and after the project's lifetime. The consortium decided not proceed this track but instead agreed upon the following:

- Data and models will be shared through the INT as CLARIN B Centre (https://centres.clarin.eu/centre/22)
- Code and scripts will be shared through our Github repository GitHub (https://github.com/signon-project, more specifically:
    - Generic: https://github.com/signon-project (public version)

    The WP-specific endpoints remain private:

    - https://github.com/signon-project-wp2
    - https://github.com/signon-project-wp3
    - https://github.com/signon-project-wp4
    - https://github.com/signon-project-wp5

This is reflected in the DMPs of individual partners.

For existing data that were used in the project, partner INT provided an sftp-site (sftp://signon@sftp.signon.ivdnt.org/private/SignLanguage) where partners could collect and share the data.

For partner Dutch Language Union there is no DMP since this partner did not use nor generate data, models or code in SignON.

3. Conclusions

D7.8 set out the principles for RDM in the SignON project and addressed the project's Open Access publication policy and how the project would meet the FAIR principles for sharing data, models and software after the lifetime of the project, how we would deal with collecting and sharing data in a GDPR compliant way during and after the lifetime of the project, data security issues and ethical aspects.

These principles became manifest in the Data Management Plans which were provided by individual SignON partners and have now been collected here in D7.10.

Annex 1: Final SignON Partners Data Management Plans

Sign Language Translation Mobile Application and Open Communications Framework

DCU Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
V1.0	Shaun O'Boyle	16/11/2021	First version of DCU DMP
V1.1	Elizabeth Mathews and Shaun O'Boyle	17/11/2021	Minor updates
V1.2	Shaun O'Boyle	06/10/2023	Updates to include additional research activities

Table of Contents

Project partners can use the DMP tools and templates provided by their own organisations as long as the guidelines outlined in this report are followed. They may also use the template that is provided by the EC[2] in the Annex section and that is included in this Appendix.

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Most data collected and generated by DCU will relate to the management, organisation, and oversight of the project, including execution and delivery in accordance with the agreed timeline. Additional data collected and generated will relate to our role in the co-creation and user response aspects of this project.

What types and formats of data will the project generate/collect?

Our data can be grouped into two categories: project coordination, and co-creation activities.

Project Coordination: No primary data will be collected or generated. Most data will be provided by project partners as summaries of their activities. The format for these data will include written and video reports and documentation.

Co-Creation Activities: Data generated by co-creation will include online surveys to seek feedback on potential use cases for the SignON application, and to evaluate the effectiveness of our co-creation events and activities with their target communities and audiences. We are a data controller for video, audio, and text generated by use case recordings in the SignON app, and we will also generate video and written documentation of selected events and activities.

Will you re-use any existing data and how?

We do not intend to re-use any existing data.

What is the origin of the data?

Data related to project coordination will come from SignON partners. Data related to co-creation will come from SignON partners, participants in co-creation events and activities, and survey respondents. Data related to use case recordings comes from users of the SignON application.

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

What is the expected size of the data?

Video data and documentation of events and activities will be approximately 20GB, and written documentation of events and activities, including evaluation data, will be approximately 1GB.

To whom might it be useful ('data utility')?

Our project coordination data will primarily be useful to SignON partners, to support the successful progression of the project. Our co-creation data will benefit SignON partners in the design of their activities, and members of the public who are interested in or impacted by the SignON project. Our use case data will benefit SignON partners in the design of the SignON application. Our evaluation data will benefit SignON partners, and may be useful to other groups interested in designing similar co-creation activities.

FAIR data

1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

Primary data produced by survey inputs will be anonymous and collected via Google Forms and Qualtrics, which does not collect metadata on respondents.

What naming conventions do you follow?

No naming conventions apply to these data.

Will search keywords be provided that optimize possibilities for re-use?

Search keywords will not be provided for documentation related to project coordination or audience surveys, as these will not be publicly available. Search keywords will be included on public videos to promote the project on YouTube.

Do you provide clear version numbers?

Version numbers will be provided where applicable.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

No metadata will be generated.

2. Making data openly accessible

How will the data be made accessible (e.g. by deposition in a repository)?

Evaluation data, including survey responses and summaries and documents relating to project coordination will be shared on Google Drive, accessible only to SignON partners. Use case recordings are stored at INT, a CLARIN B Centre..

What methods or software tools are needed to access the data?

A Google Drive account and web browser are needed to access the data.

Is documentation about the software needed to access the data included?

Documentation about the software is not required to access the data.

Is it possible to include the relevant software (e.g. in open source code)?

Not applicable.

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.

Not applicable.

Have you explored appropriate arrangements with the identified repository?

Not applicable.

If there are restrictions on use, how will access be provided?

Not applicable.

Is there a need for a data access committee?

No.

Are there well described conditions for access (i.e. a machine readable license)?

No.

How will the identity of the person accessing the data be ascertained?

Access will be restricted to Google Drive accounts of SignON partners.

3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

Use case data, and summarised evaluation data and outputs from co-creation activities will be interoperable, available in standard formats regularly used by relevant research and engagement communities. These data will be accompanied by clear guidelines on usage and license.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

This is not relevant to the data generated by DCU.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Standard and accessible vocabularies will be used to ensure interoperability.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Ontology mapping will be provided where appropriate.

4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Data that are not restricted by research ethics requirements will be licensed using the appropriate Creative Commons license, with CC0 being preferred wherever possible.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Where appropriate and permitted by the Research Ethics Committee, all data usable by third parties will be shared immediately.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Any evaluation data that is likely to be relevant to third parties, during or after the project, will be collected alongside a clear request for consent to make the data available for reuse. This is most likely to apply to future publications or guides on co-creation practise.

How long is it intended that the data remains re-usable?

The length of time that evaluation data will remain reusable will be determined by the Research Ethics Committee.

Are data quality assurance processes described?

Each data collection effort will publish its own data quality assurance processes.

Allocation of resources

What are the costs for making data FAIR in your project?

Future costs will include fees for open access publishing. There are no additional expenses for securely hosting data, and all handling of data is incorporated into staff time.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

These will be covered by costs associated with the relevant work packages.

Who will be responsible for data management in your project?

Professor Andy Way will be responsible for data management.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

All long term preservation costs have been considered, and will involve low cost or free platforms. Evaluation data will be retained for the duration of the project, or as long as is required by our Research Ethics Committee. Workshop documentation and outputs will remain publicly available as legacy content for as long as possible.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

All cloud data will be secured by Google Drive. All local data will be stored on DCU servers.

Is the data safely stored in certified repositories for long term preservation and curation?

Data will be stored on DCU servers for the appropriate length of time.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Any evaluation, survey, or co-creation activity that requires the collection of personal data will require approval from the relevant Research Ethics Committee. We do not intend to collect sensitive data. All other data will be shared under the appropriate Creative Commons license.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Informed consent for data sharing and long term preservation will be included in all use case surveys and recordings, and in all evaluation surveys and questionnaires dealing with personal data.

Sign Language Translation Mobile Application and Open Communications Framework

FINCONS Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
V1.0	Marcello Paolo Scipioni	11/11/2022	First version
V2.0	Marcello Paolo Scipioni	19/12/2022	Updates to sections related to specific contributions from Fincons
v3.0	Marcello Paolo Scipioni	27/11/2023	Updates regarding the role of INT data centre being a CLARIN centre

Table of Contents

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of the data collection is to gather resources to be used in research into optimal methods of natural language processing (NLP), machine translation (MT), sign language recognition (SLR) and automated speech recognition (ASR), 3D animation and avatar synthesis, sign language (SL) understanding, and SL linguistics for the development of a smooth communication service that uses MT to translate between SL and verbal languages and facilitates the exchange of information among deaf and hard-of-hearing (DHH) and hearing individuals.

The role of Fincons is to build software tools to enable partners to perform the needed data collection activities for building new models and updating existing ones, and not directly to collect data.

The data will be used by partners in the consortium to build new models, update and evaluate existing models, improve user experience, etc.

What types and formats of data will the project generate/collect?

In the SignON project, video, audio and text data from the DHH and hearing participants will be collected and processed. Data generated in SignON can be categorized as inference data, i.e., data to be processed in a translation job, which are deleted right after the translation has been completed, and data aimed at improving models, which are retained and processed by machine translation partners in the consortium. At Fincons, we are responsible for the SignON Orchestrator, which has the goal of transferring inference data between the SignON App and the SignON Framework where information is processed until the translation process is completed and data is deleted. Moreover, as Fincons we are building software tools to enable data collection through the Machine Learning Interface, which will be used by partners in the consortium to perform data collection activities to update models and to enable incremental update of the SignON service with new models and languages.

The project will generate and collect video data (with signers), speech recordings, text documents, and metadata about individuals as far as is needed for the research. The direct identifiers of participants will include name and email address, and possibly their device identification. These will only be used for administrative purposes whereas Indirect identifiers relevant for the research purposes are gender, age group, primary language and language use in daily life. The resulting models (acoustic models, language models, sign models) will be void of personal information and will be shared at the level of open access in standard formats.

Will you re-use any existing data and how?

At Fincons we are not planning to re-use existing data. Members of the SignON consortium will use newly collected and existing speech recordings to study the research topics mentioned above. A list of existing resources used within the project can be found in deliverable D3.1. Where possible, project members will reuse existing data, either public, or from project partners or related institutions. The appropriate permissions, based on the uses allowed, will be obtained.

What is the origin of the data?

Pre-existing datasets will be provided by project partners, public broadcasters, government information services, CLARIN data centres[3], and established data warehouses such as LDC[4], ELRA[5]. Pre-existing data sets of SignON project partners are shared under the terms of the SignON Consortium Agreement, which all partners have signed up to. Since these data will contain personal data, compliance with GDPR Art.14 will be ensured.

Data collected within the SignON project will be directly input by users through the SignON Mobile App and passed to the backend remote SignON Framework services platform which is deployed at a CLARIN B Centre (the INT data centre), where it will be processed and stored as required. In particular, data acquired through the SignOn Mobile App for translation, are temporarily stored until a request is being performed and deleted afterwards, while data acquired through the Sign On Mobile App for machine translation is specifically designed for data collection from users and data collected through this app are stored at a CLARIN B Centre (again the INT data centre, in a different physical location from those temporarily acquired through the SignOn Mobile App for translation).

What is the expected size of the data?

A list of existing resources used within the project can be found in deliverable D3.1.

Data newly collected within the project will consist of audio files (containing speech data) and video files (containing sign language data), together with textual translations and related metadata, aimed at improving existing models or to build new models. Audio and video files are supposed to be no longer than 30s each in duration, and the size of each file is expected to be around 1MB for each audio file and around 100MB for each video file (depending on the quality and resolution of mobile phones used by users uploading contributions), with negligible overhead for textual data and metadata. Data collection activities are being designed by project partners to collect several audio and video files from different speakers and signers.

[3] https://www.clarin.eu/content/overview-clarin-centres
[4] https://www.ldc.upenn.edu/
[5] http://www.elra.info/en/

To whom might it be useful ('data utility')?

Fincons does not plan to directly use collected data. The data collected will be useful to machine translation partners in the consortium to update models and create new models, and for anyone who is interested in translating between different European signed and spoken languages to facilitate the exchange of information among DHH and hearing individuals.

FAIR data

1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

All data that can legally be shared with the community, after approval by the Ethics Committees responsible, will be made available on Zenodo.org linked to the associated publications, tools, and software, or in a CLARIN data center. Open Source Software (Apache 2) produced within the project will also be made available on GitHub[6]. The resulting versions of the SignON Mobile App will be shared via project partner IVDNT being a CLARIN B data centre. Data, tools, and software will all be documented and all will have a DOI assigned by the hosting platforms.

If the original data underlying a publication cannot be made available due to lack of consent, then aggregate (and in this way anonymized) data will be made available (e.g. in the publication). In this, SignON will seek every possible strategy against re-identification.

What naming conventions do you follow?

The naming convention will reflect the contents of the respective research data (catalog number, App version, SignON API version) and the year of publication.

Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be provided to optimize possibilities for re-use. Search keywords will include the name of the project (SignOn) as well as keywords relatable to the project's subject matter such as Android, IOS, APP, etc.

[6] At https://github.com/signon-project

Do you provide clear version numbers?

Zenodo.org and CLARIN data centers enforce a clear versioning scheme, and this will be used for the versioning of data and tools.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Zenodo.org and CLARIN data centers have provisions for assigning metadata. In our metadata schemes, we will make clear which versions of the App, device operating system (OS), user interface (UI) default settings and SignON Framework API versions.

2. Making data openly accessible

Which data produced and/or used in the project will 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.

SignON investigates sign language (SL), speech and text from DHH and hearing people. Both at the European and at a national level, privacy regulations require that researchers secure ethical approval and informed consent before the publication of data from human participants is permitted. As a result, participant data can only be shared if informed consent to share the data was given by the data subjects.

How will the data be made accessible (e.g. by deposition in a repository)?

All data that can be collected without the approval of an Ethics Committee will be made openly available from Zenodo or CLARIN and indexed in OpenAIRE. For the other data, informed consent as approved by the Ethical Committee will be taken for making data accessible.

Data will be of varying natures and published in commonly used, standard formats. All data will be accompanied by documentation of how to read and use it. If necessary, the required software tools will be described or included.

For educational purposes, the project records presentations of workshops and webinars when the presenter agrees to record. These recordings will be made fully public on Zenodo.org after curation and signed approval by the presenters.

What methods or software tools are needed to access the data?

All data and publications will be stored on Zenodo.org, in CLARIN data centres which are supported by OpenAIRE and H2020.

Is documentation about the software needed to access the data included?

All data will be accompanied with documentation of how to read and use it.

Is it possible to include the relevant software (e.g. in open source code)?

If necessary, the required software tools will be described or included.

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.

The SignON Open Source Software produced within the project will be made available on GitHub[7].
All data and publications will be stored on Zenodo.org, in CLARIN data centers which are supported by OpenAIRE and H2020.
All above platforms also have provisions for assigning metadata.

Have you explored appropriate arrangements with the identified repository?

SignON will store its data in CLARIN data centres which are supported by OpenAIRE and H2020.

If there are restrictions on use, how will access be provided?

During the project and limited to SignON project partners a Data Transfer Agreement is established to exchange such data.

Is there a need for a data access committee?

Data that is not openly accessible will be available on-site, or using secure remote access, to individual researchers after approval by the relevant Ethics Committees. Data access will be decided by the Ethics Committees of the participating institutions. Decisions will be made on a case-by-case basis by the

[7] https://github.com/signon-project

partners involved for data collections generated by the project. No SignON data access committee is needed.

Are there well described conditions for access (i.e. a machine readable license)?

Zenodo.org provides well-described conditions for access[8]. For data that is not openly accessible a DUA (Data User Agreements) will apply.

How will the identity of the person accessing the data be ascertained?

Users are required to register to use the repository.

3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All data will be published in the formats commonly used in the research communities concerned. If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are open access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no open access options, this data format too will be made available alongside the open data format.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Metadata like gender, age interval, hearing status, will be associated with newly recorded data.
For data that will be hosted in a CLARIN data center, the format requirements of the CLARIN data center will be followed to make data interoperable[9].

[8] See http://about.zenodo.org/policies/
[9] see https://www.clarin.eu/content/interoperability

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

If available, standard vocabularies for all data types will be used whenever possible to ensure inter-disciplinary interoperability and re-use.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Data, text, and software will be published under Creative Commons or, for software, Apache 2 license[10], unless there are contractual or legal reasons that make this not possible. For data that is not openly accessible a DUA (Data User Agreements) will apply.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be made available after the main publication based on the data having been published, or earlier, if possible, but not longer than 6 months after completion and publication of the data. If an embargo is sought beyond these times, the reasons and duration for the embargo will be given.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions, or is prohibited by laws or regulations in the countries of the participants.

[10] This will be described in D6.7 “First SignON Sustainable Exploitation, Innovation and IPR Plans”

How long is it intended that the data remains re-usable?

Data available on Zenodo.org and CLARIN data centres will remain available without a time limit. In order to share data after the lifetime of the project, consent of the data subjects will be sought by asking permission for re-use of the data for the following purposes:

- Train and test artificial intelligence systems for sign language recognition
- Train and test artificial intelligence systems for sign language synthesis
- Train and test artificial intelligence systems for sign language translation
- Train and test artificial intelligence systems for sign language understanding
- Translation studies from and to sign languages
- Linguistic studies about the properties of sign languages, spoken languages or written texts
- Train and test artificial intelligence systems for text-to-text translation, i.e. machine translation
- Train and test systems for generation of a virtual signer, i.e. a 3D avatar
- Train and test artificial intelligence systems for spoken language recognition or synthesis (audio modality)

This will be included in the informed consent forms.

Are data quality assurance processes described?

Yes, the data quality assurance process is described. The quality assurance processes will include the provision of results along with the data and the peer-review of publications based on the data.

Allocation of resources

What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting of the data generated by the project.

The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo or CLARIN as repositories.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

There are no extra costs to SignON consortium partners for making the data FAIR. The costs, in time and effort, to upload data and publications to Zenodo or CLARIN Data centres are marginal and covered by the project and its overhead provisions.

Who will be responsible for data management in your project?

The Principal Investigators (PIs) of the project from the different institutions will be responsible for data management, including making data and publications FAIR. The Coordinator will oversee the implementation. All publications and data that can be published will be stored at Zenodo or CLARIN. Data that cannot be published or has other restrictions will be stored at Fincons or on secure remote storage. Only descriptions and contact information of this latter data will be published to make them findable.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

At the project level, we have made provisions for long-term secure storage of data and publications in the form of repositories. Data that is uploaded to Zenodo.org and CLARIN data centers will be available without a time limit. Data that is not open will be stored according to Fincons standard procedures. The use of the long-term repositories does not constitute a cost for the SignON project.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The recordings and the associated metadata will be transferred from participant's mobile phones to a locally secured server of one of the project partners (INT). They will be stored on a bucket on the Object Storage system which is physically located on a disk in the infrastructure of SignON partner INT. To access such data SignON partners with permission will receive MinIO credentials hosted on the infrastructure from INT. These partners are listed in the application. The SignON recording app is GDPR compliant in storing personal data.

As participant information will be erased, after the recordings they will not be further informed (on an individual basis) of the outcomes of the project, but we will point the participants to our website as information about project outputs will be made available via these media.

Under GDPR, data subjects have the right to withdraw consent, and consent must be as easy to withdraw as it was to give. We will therefore ensure that a mechanism is in place for tracking and managing consents, and any project contingencies which may be required as a result of consent being withdrawn. On the other hand, it must be stressed that data subjects cannot request the deletion of the collected data (GDPR Art. 17 3 (d)).

For exchanging data within the project, the servers of project partners IVDNT (which is a CLARIN B centre) will be used, and for which SignON has created its own group account. Data exchange between project partners will follow guidelines laid out in the templates for the project's Data Transfer Agreements (D7.9) and will use encrypted file transfer facilities such as FileSender. A template for Data Transfer Agreements for joint controllers concerning data obtained from third parties such as broadcast companies has also been provided in D7.9.

Is the data safely stored in certified repositories for long term preservation and curation?

Zenodo.org and CLARIN data centers have adequate provisions for data security.

Data protection support is provided by the DCU Data Protection Officer, Martin Ward, and the DCU Data Protection Coordinator, Joan O'Connell. They can be contacted at data.protection@dcu.ie. For queries around the protection of personal data with respect to the SignON project, please contact signon-data-protection@adaptcentre.ie.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are strong legal and ethical limitations on the access of personal data. All other data will be shared under Creative Commons or open-source licenses. If possible, informed consent for the long-term preservation and sharing of the data will be sought from the data subjects. In agreement with GDPR, all participants in any data collection process will have access to information in plain language and/or a signed language, as is their preference. The consortium takes the position that special data protection provisions above those required for personal data are not needed in this project.

As discussed above in DMP, we are not collecting any metadata that relates to the health of any individual - our focus is solely on participants' linguistic preferences and their experiences with and attitudes to MT (Machine Translation) and the SignON App**.**

For the use case recordings in the hospitality domain that were carried out within the project, participants will be asked for consent to publish the data as a corpus for re-use outside the project after the project end.

The ethics application for the use case recordings is filed at the REC of project coordinator DCU.

Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case. The corpus will be made available through a CLARIN data center.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, informed consent for data sharing and long-term preservation is included in questionnaires dealing with personal data. New data that will be collected will require a two-stage process of ethics clearance. That is, when we seek to collect data (e.g., from participants in focus groups, or in creating additional data sets to supplement existing corpora), partners will prepare an ethics application for their home institution, or, if they are a non-university partner, an application for research ethics approval will be submitted via the coordinating partner institution, Dublin City University. Before submitting their application to their institutional research ethics committee, the application will be reviewed by the SignON Ethics Committee.

Sign Language Translation Mobile Application and Open Communications Framework

INT Data Management Plan 2023

Authors: Vincent Vandeghinste and Bob Boelhouwer

Version #	Implemented by	Revision Date	Description of changes
V0.1	Vincent Vandeghinste	<dd/mm/yyyy>	<…..>
V0.2	Bob Boelhouwer	<15/11/2022>	No changes apart from some spelling
V0.3	Bob Boelhouwer	05/10/2023	No changes

Table of Contents

1. Data Summary

1.1. What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data that INT will collect concerns existing data, either from project partners in the SignON project, or from external sources. The goal of this data collection is redistribution, primarily within the SignON consortium, but also to external partners, if licenses and informed consents permit.

The purpose of the data collection is to gather resources to be used in research into optimal methods of natural language processing (NLP) and automated speech recognition (ASR) for the development of a smooth communication service that uses MT to translate between verbal languages and facilitates the exchange of information among deaf and hard-of-hearing (DHH) and hearing individuals. That is, the data will be used to build new models, update and evaluate existing models, improve user experience, etc. Models and algorithms that will exploit the data will be designed and developed such that the information contained therein cannot be deduced to any individual.

1.2. What types and formats of data will the project generate/collect?

We expect to collect different types of data. Concerning Sign Language (SL) users, we will collect video material that has the SL user at its focus, and in which SL can be considered the source language. Additionally, we also collect video material of public broadcasts which have an SL interpreter in a section of the screen. In the latter case, the video material also contains audio and possibly subtitles and autocue. We also collect existing audio material, both from *regular* speakers as from *atypical* speakers. These audiovisual data will be complemented with text collections, which will again consist of existing text corpora.

1.3. Will you re-use any existing data and how?

All datasets that will be collected and redistributed through INT will consist of existing data or new data created by partners in the SignON Project. No primary data will be created by INT.

1.4. What is the origin of the data?

Datasets originate from different sources. Concerning the different Sign Language corpora, which are normally only available through online interfaces we have received the data in an easily downloadable format from the respective corpus creators. This concerns the Corpus Vlaamse Gebarentaal https://www.corpusvgt.be/, the Corpus Nederlandse Gebarentaal https://www.corpusngt.nl/, and the Signs of Ireland Corpus.

The Belgian VGT Corona corpus consists of the videos available at https://news.belgium.be/nl/corona. We also aim to collect data from VRT, the Flemish public broadcaster.

The Content4All Newscorpus containing data with VGT, at https://www.cvssp.org/data/c4a-news-corpus/ were made available by the University of Surrey.

Other datasets are downloaded from the CLARIN Virtual Language Observatory at https://vlo.clarin.eu, or from other public sources.

Furthermore, use case recordings and the associated metadata created by partners within the SignON Project will be stored on the locally secured server of INT. We can distinguish the hospitality recordings that were made with the ML app and the HoReCo test recordings.

1.5. What is the expected size of the data?

As it concerns video data, the expected size of the data is considerable. At the end of the project, the size of the data was 3.85 terabyte.

1.6. To whom might it be useful ('data utility')?

The data is primarily collected for NLP and AI researchers within the SignON project. When licenses and informed consents permit we make datasets as widely available as allowed.

FAIR data

1. Making data findable, including provisions for metadata

Data sets that are shared within the consortium will be accessible via an ftp connection. The data will be organized in a traditional hierarchical tree-structure with meaningful names for the path elements. At the root level we will provide a document that will list all data sets and the location within the tree structure.

Since the amount of data sets and the number of users are limited it will not be necessary to implement more sophisticated search and discovery methods.

Data sets that are not yet available for a wider audience will be made available in the “Taalmaterialen” repository of language resources of the INT (“https://taalmaterialen.ivdnt.org/”) provided that the owners of the data sets agree and that the data sets are cleared for personal data.

2. Making data openly accessible

Since the INT does not own or produce any of the data that is used within SignOn, the decision on making the data openly accessible lies with the owners.

The INT offers the possibility to accommodate the resources in the “Taalmaterialen” repository.

The data will be stored in a secure, reliable and certified long-term preservation repository. Users will be able to download the data from the Taalmaterialen website using a standard browser. Users are required to register and agree with licensing conditions before they are able to download any data. Metadata will be published through the CLARIN Virtual Language Observatory (VLO). The data will also be identifiable and locatable by means of a PID.

With respect to the use case recordings: Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case. The corpus will provisionally be made available through INT being a CLARIN data center,and a SignON project partner.

3. Making data interoperable

The general policy of the INT is to publish data in formats that follow free and open standards.

4. Increase data re-use (through clarifying licences)

Since the INT is not the owner of the data that we provide to others, the decision on what license to apply to the data is up to the owners. The INT only distributes data that is free of charge.

Allocation of resources

The Taalmaterialen repository in which we intend to distribute data from Signon is a standing service of the INT. No extra expenses will be required to make the data available through this service.

INT WP3 person months are dedicated to data collection and redistribution of existing data sets

Data security

The data that is available for internal use within the SigOn project is only accessible via a login procedure using encrypted scp and sftp.

The data is stored on a bucket on the Object Storage system (MinIO, an open source solution compatible with the Amazon S3 protocol) which is physically located on a disk in the infrastructure of SignON partner INT. To access such data SignON partners with permission will receive MinIO credentials hosted on the infrastructure from INT. These partners are listed in the application. The SignON recording app is GDPR compliant in storing personal data (this is part of the requirements of the EC for funding the SignON project).

Ethical aspects

Since a number of data sets contain footage of signers, we will verify that the owners and providers of the data hold the appropriate consents before making them available to partners with the SignOn Project.

For the use case recordings in the hospitality domain that were carried out within the project, participants will be asked for consent to publish the data as a corpus for re-use outside the project after the project end.

The ethics application for the use case recordings is filed at the REC of project coordinator DCU.

As participant information will be erased after the recordings they will not be further informed (on an individual basis) of the outcomes of the project, but we will point the participants to our website as information about project outputs will be made available via these media.

Sign Language Translation Mobile Application and Open Communications Framework

UPV/EHU Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
V1.0	Gorka Labaka	02/11/2022	First version

Table of Contents

1. Data Summary

1.1. What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data that UPV/EHU will collect concerns existing data, either from project partners in the SignON project, or from external sources. The goal of this data collection is redistribution, primarily within the SignON consortium, but also to external partners, if licences and informed consents permit.

The data collection is aimed at researching automatic spoken language and sign language translation systems. That is, in order to build and to prepare models that are capable of translating from and to sign and spoken languages, parallel data is required. This type of model will help people, especially deaf and hard of hearing (DHH) people, in their daily lives to communicate.

1.2. What types and formats of data will the project generate/collect?

The collected data will be in text format for spoken languages (in txt or json files) and in video format (mp4, for example) for sign language data (except for glosses, which will also be in text format similar to spoken language data).

1.3. Will you re-use any existing data and how?

All datasets that will be collected and redistributed through UPV/EHU will consist of existing data. No primary data will be created, but existing datasets may be complemented with annotations or translations.

1.4. What is the origin of the data?

The data we gather can be publicly available (open source datasets) and is usually shared with the following constraints: citing the owners, not using it for commercial purposes and sharing it under the same terms.

Pre-existing data sets of SignON project partners are shared under the terms of the SignON Consortium Agreement, which all partners have signed up to. Since these data will contain personal data, compliance with GDPR Art.14 will be ensured.

1.5. What is the expected size of the data?

Datasets may vary from a few megabytes to several gigabytes. For spoken languages more data is available and, hence, the size of those datasets is in the upper bound. However, sign language datasets are often smaller and may be close to the lower bound, depending on the dataset.

1.6. To whom might it be useful ('data utility')?

The data is primarily collected for NLP and AI researchers within the SignON project. When licences and informed consents permit it, we will make datasets as widely available as possible.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

2.1.1. Are the data produced and/or used 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)?

The UPV/EHU will not produce spoken or sign language data. However, whenever data from partners is used, this must be approved by the responsible Ethics Committees. Any Free Libre Open Source Software produced within the project will be made available on GitHub (https://github.com/signon-project). The models generated as output will be stored and shared in GitHub and in the CLARIN infrastructure provided by partner IVNT as a CLARIN B Centre.

2.1.2. What naming conventions do you follow?

The naming convention involves including all the necessary information to identify the data such as the original name, date of publication, purpose and so on.

2.1.3. Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be available for optimised searches, including the name of the project, work package (if applicable) and other keywords related to the project's research.

2.1.4. Do you provide clear version numbers?

The code in GitHub is organised using tags and dates of publication of the code. These tags include keywords such as the work package, the associated deliverable or the new functionality added to identify them.

2.1.5. What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

The software developed has a README file that explains the content of the repository, the data structure that needs to be followed and the necessary steps to make the software work. All the collected data have files attached that specify the authorship, date of creation, description and other related details.

2.2. Making data openly accessible

2.2.1. Which data produced and/or used in the project will 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.

Data collected from the internet is made available by their owners. In case data shared within the consortium requires informed consent, the data will not be shared unless the subjects give their permission. Regarding the models developed by UPV/EHU, only the models that are considered shareable will be published.

2.2.2. How will the data be made accessible (e.g. by deposition in a repository)?

All data that can be collected without the approval of an Ethics Committee will be made openly available from Zenodo or CLARIN and indexed in OpenAIRE. For the other data, informed consent as approved by the Ethics Committee will be taken for making data accessible.

Data will be of varying natures and published in commonly used, standard formats. All data will be accompanied by documentation of how to read and use it. If necessary, the required software tools will be described or included.

For educational purposes, the project records presentations of workshops and webinars when the presenter agrees to record. These recordings will be made fully public on Zenodo.org after curation and signed approval by the presenters.

2.2.3. What methods or software tools are needed to access the data?

All open data and publications will be stored on Zenodo.org, in CLARIN data centers (of project partner IVDNT) which are supported by OpenAIRE and H2020. If specific methods or software to access the data are required, they will be specified.

2.2.4. Is documentation about the software needed to access the data included?

No additional documentation is needed to open the research data.

2.2.5. Is it possible to include the relevant software (e.g. in open source code)?

The code that is shareable will be open source and will be published in public repositories (GitHub).

2.2.6. 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.

All the code will be available in GitHub, the models related to the project will be stored in a common secure space within the IVDNT CLARIN infrastructure (shared among partners) and experimental results and models that need and can be shared with a wider community will be shared via GitHub.

All data and publications will be stored on Zenodo.org, in CLARIN data centers which are supported by OpenAIRE and H2020.

All above platforms also have provisions for assigning metadata.

2.2.7. Have you explored appropriate arrangements with the identified repository?

The project has already set-up the repositories.

2.2.8. If there are restrictions on use, how will access be provided?

During the project and limited to SignON project partners a Data Transfer Agreement is established to exchange such data.

2.2.9. Is there a need for a data access committee?

UPV/EHU will not produce any of the data used within the consortium and, hence, does not require an access committee.

2.2.10. Are there well described conditions for access (i.e. a machine readable license)?

Yes, Zenodo.org provides well-described conditions for access[11]. For data that is not openly accessible a DUA (Data User Agreements) will apply (see above).

2.2.11. How will the identity of the person accessing the data be ascertained?

Users are required to be registered and logged in to use the repositories.

2.3. Making data interoperable

2.3.1. Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All data will be published in the formats commonly used in the research communities concerned. If available, public guidelines for metadata vocabularies, standards or methodologies will be followed. Standard data formats for which there are open access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no open access options, this data format too will be made available alongside the open data format.

2.3.2. What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that will be hosted in a CLARIN data center, the format requirements of the CLARIN data center will be followed to make data interoperable[12].

2.3.3. Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

If available, standard vocabularies for all data types will be used whenever possible to ensure inter-disciplinary interoperability and re-use.

[11] See http://about.zenodo.org/policies/

[12] see https://www.clarin.eu/content/interoperability

2.3.4. In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

2. 4. Increase data re-use (through clarifying licences)

2.4.1. How will the data be licensed to permit the widest re-use possible?

UPV/EHU is not the owner of the data used within the consortium. The decision of what license should be used lies in the owners of the data.

2.4.2. When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

As the UPV/EHU does not generate its own data, the decision of making the data available should be taken by the owners of the data.

2.4.3. Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions or is prohibited by laws or regulations in the countries of the participants.

2.4.4. How long is it intended that the data remains re-usable?

Data available on Zenodo.org and CLARIN data centres will remain available without a time limit. The UPV/EHU does not produce sensitive data. Taking the appropriate measures to share the data after the lifetime of the project should be the responsibility of the owners of the data.

2.4.5. Are data quality assurance processes described?

This decision is up to the owners of the data as the UPV/EHU will not generate new data.

3. Allocation of resources

3.1. What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of curating and formatting the data collected in the project. There are no expenses for hosting the data as the servers of the UPV/EHU will be used. No new data will be generated, so there won't be associated costs.

The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using the Zenodo repository.

3.2. How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

As such, there are no extra costs for making the data FAIR. The costs, in time and effort, to upload data and publications to Zenodo.org or CLARIN data centres are marginal and covered by the project and its overhead provisions.

3.3. Who will be responsible for data management in your project?

The Principal Investigators (PIs) of the project will be responsible for the data management (with the coordinator overseeing it).

3.4. Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Data will be stored in the servers of the UPV/EHU. There is not an associated cost to it. The data will be kept for as long as it is necessary.

4. Data security

4.1. What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Data and models will be stored in servers of the UPV/EHU. These require to be logged in order to access them and automatically create backup copies every day. Hence, data is safely stored. Models will also be shared within the consortium using the IVDNT CLARIN infrastructure and GitHub. The hosting of both platforms will be in charge of the data security. Published data will be deposited in Zenodo.org or CLARIN data centres for long-term preservation and curation.

4.2. Is the data safely stored in certified repositories for long term preservation and curation?

Zenodo.org and CLARIN data centres have adequate provisions for data security.

Apart from that, Data protection support is provided by the DCU Data Protection Officer, Martin Ward, and the DCU Data Protection Coordinator, Joan O'Connell. They can be contacted at data.protection@dcu.ie.

For queries around the protection of personal data with respect to the SignON project, the queries can be raised at signon-data-protection@adaptcentre.ie

5. Ethical aspects

5.1. Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are strong legal and ethical limitations on the access of personal data. All other data will be shared under Creative Commons or open-source licenses. If possible, informed consent for the long-term preservation and sharing of the data will be sought from the data subjects. As the UPV/EHU does not produce any sensitive data, our only constraint is that the data we receive from partners should have taken this into consideration.

5.2. Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

UPV/EHU will not use questionnaires as it will not be involved in the collection of personal data.

Sign Language Translation Mobile Application and Open Communications Framework

MAC Data Management Plan 2023

Authors: John O’Flaherty, Ed Keane, Connor O’Reilly

Version #	Implemented by	Revision Date	Description of changes
V2.0	John O'Flaherty	04/12/2023	MAC DMP 2023 updated with feedback from the SignON research Ethics Committee
V1.0	John O'Flaherty	05/10/2023	Update of the MAC 2022 DMP - no significant changes to the content were made.

Table of Contents

The SignON project has received funding from the European Union's Horizon 2020 Programme under Grant Agreement No. 101017255. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the SignON project or the European Commission. The European Commission is not liable for any use that may be made of the information contained therein.

The Members of the SignON Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the SignON Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

1. Data Summary

1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of MAC's data collection in the SignON project is to develop the SignON mobile applications:

1. **SignON SLMT (Sign Language Machine Translation) App** - that translates between different European signed and spoken languages to facilitate the exchange of information among deaf and hard-of-hearing (DHH) and hearing individuals.
2. **SignON ML (Machine Learning) App** - that enables users to record Sign Language (SL) video or speech audio clips for storing on the SignON platform, to train and improve the SIgnON SL and speech Machine Translation (MT) systems,

The SignON applications have lightweight software and minimal transitory data running on a standard mobile device, and interact with the cloud-based distributed SignON Framework platform server computationally heavy and data-intensive services.

1.2 What types and formats of data will the project generate/collect?

In the SignON project, video, audio and text data from the DHH and hearing participants is input and output to users by the SignON mobile applications and processed in the SignON backend platform, as described in section 2 of D7.10.

The required data for the SignON mobile applications, listed in section 1.1, is directly input by users and passed to the backend remote SignON Framework services platform, where it is processed and stored as required for research purposes, as described in section 3 of D7.10. While data for the applications' output come from the SignON platform for presentation to users. So the applications' data is temporarily stored on the user's own mobile device.

1.3 Will you re-use any existing data and how?

At MAC we are responsible for the SignON mobile applications listed in section 1.1. These output/input transitory data directly to/from users, running on a standard mobile device, and interact with the cloud-based SignON Framework platform video/audio/text data-intensive services, as described in section 1.2.

1.4 What is the origin of the data?

The required data for operation and use of the SignON mobile applications, is temporarily stored while being directly input by and output to users on their own mobile devices as described in section 1.2.

.

Pre-existing data sets of SignON project partners are shared under the terms of the SignON Consortium Agreement, which all partners have signed up to. Since these data contain personal data, compliance with GDPR Art.14 is ensured.

1.5 What is the expected size of the data?

The size of the SignON Mobile Apps research data is minimal – less than 100MB for any mobile device.

1.6 To whom might it be useful ('data utility')?

The SignON Framework data input/output process of the SignON applications, listed in section 1.1, will be useful for users and researchers interested in translating between different European signed and spoken languages to facilitate the exchange of information among DHH and hearing individuals. The transitory data itself of the SignON applications in each user's phone is used to enable them to interact with the cloud-based SignON Framework platform video/audio/text data-intensive services, so it's not useful beyond that.

2. FAIR data

2.1 Making data findable, including provisions for metadata

2.1.1 Are the data produced and/or used 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)?

All of MAC's SIgnON applications' data adhere to section 3 of D7.10.

All data that can legally be shared with the community, after approval by the Ethics Committees responsible, will be made available on Zenodo.org linked to the associated publications, tools, and software, or in the CLARIN data center. Any Open Source Software (Apache 2) produced within the project will also be made available on GitHub[13]. The resulting versions of the SignON mobile applications,

[13] At signon-project-wp2 (github.com)

listed in section 1.1, will be shared via IVDNT as CLARIN B Centre[14][15]. Data, tools, and software will all be documented and all will have a DOI assigned by the hosting platforms.

If the original data underlying a publication cannot be made available due to lack of consent, then aggregate (and in this way anonymized) data will be made available (e.g. in the publication). In this, SignON will seek every possible strategy against re-identification.

2.1.2 What naming conventions do you follow?

The naming convention reflects the contents of the respective research data (catalog number, App version, SignON API version) and the year of publication.

2.1.3 Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be provided to optimize possibilities for re-use. Search keywords will include the name of the project (SignOn) as well as keywords relatable to the project's subject matter such as Android, IOS, APP, etc.

2.1.4 Do you provide clear version numbers?

Zenodo.org and CLARIN data centers enforce a clear versioning scheme, and this will be used for the versioning of data and tools.

2.1.5 What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Zenodo.org and CLARIN data centers have provisions for assigning metadata. In our metadata schemes, we make clear which versions of each SignON application, device operating system (OS), user interface (UI) default settings and SignON Framework API versions.

[14] https://centres.clarin.eu/centre/22

[15]

2.2 Making data openly accessible

2.2.1 Which data produced and/or used in the project will 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.

As described in section 3.2 of D7.10, SignON investigates sign language (SL), speech and text from DHH and hearing people. Both at the European and at a national level, privacy regulations require that researchers secure ethical approval and informed consent before the publication of data from human participants is permitted. As a result, participant data can only be shared if informed consent to share the data was given by the data subjects. This will always be on a restricted access basis for any relevant data that we collect for this project at MAC.

2.2.2 How will the data be made accessible (e.g. by deposition in a repository)?

All data that can be collected without the approval of an Ethics Committee will be made openly available from Zenodo or CLARIN and indexed in OpenAIRE. As per section 3.2 of D7.10. For the other data, informed consent as approved by the Ethical Committee will be taken for making data accessible.

Data will be of varying natures and published in commonly used, standard formats. All data will be accompanied by documentation of how to read and use it. If necessary, the required software tools will be described or included.

For educational purposes, the project records presentations of workshops and webinars when the presenter agrees to record. These recordings will be made fully public on Zenodo.org after curation and signed approval by the presenters.

2.2.3 What methods or software tools are needed to access the data?

All open data and publications will be stored on Zenodo.org and in CLARIN data centers which are supported by OpenAIRE and H2020. Versions of the SignON applications, listed in section 1.1, that are developed within the scope of the project are stored in the SignON WP2 GitHub repository[16]. Experimental results and Apps' versions that need and can be shared with the wider community will be shared through the IVDNT as CLARIN B Centre[17].

[16] https://github.com/signon-project-wp2 and https://github.com/SignON-project
[17]https://centres.clarin.eu/centre/22

2.2.4 Is documentation about the software needed to access the data included?

No additional documentation is needed to open the SignON application's research data.

2.2.5 Is it possible to include the relevant software (e.g. in open source code)?

Yes.

2.2.6 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.

The SignON applications' Open Source Software (Apache 2) produced within the project will be made available on GitHub[18]

All data and publications will be stored on Zenodo.org and in the CLARIN data center which are supported by OpenAIRE and H2020.

Experimental results and SignON applications' versions that need and can be shared with the wider community will be shared via IVDNT as CLARIN B Centre

All above platforms also have provisions for assigning metadata.

2.2.7 Have you explored appropriate arrangements with the identified repository?

SignON has its own group account at CLARIN B Centre and https://github.com/signon-project on GitHub.

2.2.8 If there are restrictions on use, how will access be provided?

During the project and limited to SignON project partners the D7.9 Data Transfer Agreement is established to exchange such data.

2.2.9 Is there a need for a data access committee?

For data that is not openly accessible, restricted access limitations will be formulated in user licenses DUA (Data User Agreements). Requests to access such data outside the SignON project will go via the MAC SignON R&D Team Manager and will require acceptance of the DUA.

2.2.10 Are there well described conditions for access (i.e. a machine readable license)?

[18] signon-project-wp2 (github.com)

Yes, Zenodo.org provides well-described conditions for access [19]. For data that is not openly accessible a DUA (Data User Agreements) will apply (see above).

2.2.11 How will the identity of the person accessing the data be ascertained?

Users are required to register to use the repository.

2.3 Making data interoperable

2.3.1 Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All data will be published in the formats commonly used in the research communities concerned, as described in section 3.3 of D7.10. If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are open access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no open access options, this data format too will be made available alongside the open data format.

2.3.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that will be hosted in a CLARIN data center, the format requirements of the CLARIN data center will be followed to make data interoperable[20].

2.3.3 Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

If available, standard vocabularies for all data types are used whenever possible to ensure inter-disciplinary interoperability and re-use.

2.3.4 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

[19] See http://about.zenodo.org/policies/
[20] see https://www.clarin.eu/content/interoperability

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

2.4 Increase data re-use (through clarifying licenses)

2.4.1 How will the data be licensed to permit the widest re-use possible?

All data and text is published under Creative Commons, and for software, Apache 2 license[21], unless there are contractual or legal reasons that make this not possible.

For data that is not openly accessible, restrictions mentioned under 2.2.8 and 2.2.9 will apply.

2.4.2 When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

As described in section 3.4 of D7.10, data will be made available after the main publication based on the data having been published, or earlier, if possible, but not longer than 6 months after completion and publication of the data. If an embargo is sought beyond these times, the reasons and duration for the embargo will be given.

2.4.3 Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions, or is prohibited by laws or regulations in the countries of the participants.

2.4.4 How long is it intended that the data remains re-usable?

Data available on Zenodo.org and CLARIN data centers will remain available without a time limit.

To share data after the lifetime of the project, consent of the data subjects will be sought by asking permission for re-use of the data for the following purposes:

- Linguistic studies about the properties of sign languages, spoken languages, or written texts

[21] This is described in D6.7 “First SignON Sustainable Exploitation, Innovation and IPR Plans”

- Train and test artificial intelligence systems for text-to-text translation, i.e. machine translation
- Train and test artificial intelligence systems for spoken language recognition or synthesis (audio modality)

This will be included in the informed consent forms.

2.4.5 Are data quality assurance processes described?

Yes, the data quality assurance process is described. The quality assurance processes will include the provision of results along with the data and the peer-review of publications based on the data.

3. Allocation of resources

3.1 What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting the data generated by the project. The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo and the CLARIN B Centre as repositories.

3.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

As such, there are no extra costs for making the data FAIR. The costs, in time and effort, to upload data and publications to Zenodo.org or CLARIN Data centers are marginal and covered by the project and its overhead provisions.

3.3 Who will be responsible for data management in your project?

The MAC R&D Team Manager is responsible for data management, including making data and publications FAIR.

The coordinator will oversee the implementation, as described in section 5 of D7.10. All publications and data that can be published will be stored at Zenodo and/or CLARIN B Centre. Data that cannot be published or has other restrictions will be stored at MAC as the hosting institution. Only descriptions and contact information of this latter data will be published to make them findable.

3.4 Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

At the project level, we have made provisions for long-term secure storage of data and publications in the form of repositories (see section 5 of D7.10). Data that is uploaded to Zenodo.org and CLARIN data centers will be available without a time limit. Data that is not open is being stored according to MAC's standard procedures. The use of the long-term repositories does not constitute a cost for the SignON project. The minimum period for data retainment is set to 5 years at MAC.

4. Data security

4.1 What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Under GDPR, data subjects have the right to withdraw consent, and that consent must be as easy to withdraw as it was to give. We therefore ensure that consents are properly managed and can be withdrawn at any time. On the other hand, it must be stressed that if deletion is likely to render impossible or seriously impair the achievement of the objectives of the processing, then data subjects cannot request the deletion of the collected data according to GDPR Art. 17 3 (d)[22].

Unpublished data will be stored in MAC, which has its own data security provisions. The data will be stored for at least 5 years for instance, for reasons of commercial or scientific integrity. Published data will be deposited in Zenodo.org or CLARIN data centers for long-term preservation and curation.

For exchanging versions of the SignON applications of section 1.1, and data within the project, the SignON GitHub Repository[23] used. Data exchange between project partners follows guidelines laid out in the templates for the project's Data Transfer Agreements (D7.9) and uses encrypted file transfer facilities such as FileSender[24]. A template for Data Transfer Agreements for joint controllers concerning data obtained from third parties such as broadcast companies has also been provided in D7.9.

4.2 Is the data safely stored in certified repositories for long term preservation and curation?

Zenodo.org and CLARIN data centers have adequate provisions for data security.

Apart from that, as described in section 6 of D7.10, Data protection support is provided by the DCU Data Protection Officer, Martin Ward, and the DCU Data Protection Coordinator, Joan O'Connell. They can be contacted at data.protection@dcu.ie.

For queries around the protection of personal data with respect to the SignON project, the queries can be raised at signon-data-protection@adaptcentre.ie

[22] Art. 17 GDPR – Right to erasure ('right to be forgotten') - General Data Protection Regulation
[23] At https://github.com/signon-project
[24] SURFfilesender: send large files securely and encrypted

5. Ethical aspects

5.1 Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are strong legal and ethical limitations on the access of personal data, as discussed in section 7 of D7.10 All other data will be shared under Creative Commons or open-source licenses. If possible, informed consent for the long-term preservation and sharing of the data will be sought from the data subjects. In agreement with GDPR, all participants in any data collection process will have access to information in plain language and/or a signed language, as is their preference. The consortium takes the position that special data protection provisions above those required for personal data are not needed in this project.

As discussed in section 1, MAC is not collecting any metadata that relates to the health of any individual - our focus is solely on participants' linguistic preferences and their experiences with and attitudes to MT (Machine Translation) and the SignON applications.

5.2 Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, informed consent for data sharing and long-term preservation is included in questionnaires dealing with personal data.

As discussed throughout D7.10, new data that will be collected will require a two-stage process of ethics clearance. That is, when we seek to collect data (e.g. from participants in focus groups, or in creating additional data sets to supplement existing corpora), partners will prepare an ethics application for their home institution, or, if they are a non-university partner (such as MAC), an application for research ethics approval will be submitted via the coordinating partner institution, Dublin City University. Before submitting their application to their institutional research ethics committee, the application will be reviewed by the SignON Ethics Committee.

6. Other Issues

6.1 Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No, as any SignON applications' research data that MAC will collect will not require such. This means that data collected from data subjects will be stored and managed by MAC.

Sign Language Translation Mobile Application and Open Communications Framework

Universitat Pompeu Fabra Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
~~<V0.1>~~	~~Santiago Egea Gómez~~	~~04/11/2022~~	~~Review of first draft~~
v.0.1	Josep Blat	24/10/2023	Revision of 2022 plan with the perspective of UPF-GTI
v.0.2	Santiago Egea	24/10/2023	DMP update for UPF-TALN

Table of Contents

Project partners can use the DMP tools and templates provided by their own organizations as long as the guidelines outlined in this report are followed. They may also use the template that is provided by the EC[25] in the Annex section and that is included in this Appendix.

Data Summary

UPF takes part in SignON through two different research groups, with different aims. UPF-GTI deals mainly with the synthesis of the virtual signer (aka signing avatar) while UPF-TALN is involved in computational linguistic tasks.

What is the purpose of the data collection/generation and its relation to the objectives of the project?

UPF-GTI intends to

- collect *animations* of signs, to give more realism to the signing of the virtual signer
- develop *software* that allows crowd-sourced collection of sign animations
- extend the Behaviour Markup Language (BML) *specification* to better support conversations involving virtual signers; throughout the project, supporting other specifications, such as SiGML, which have been used to represent collections of signs has been included as a goal
- develop other software useful for the synthesis of virtual signers

UPF-TALN

Linguistic models based on Neural Networks demand a high quantity of data on which models will learn to solve the specific task. The success of these Machine Translation models highly depends on the quantity and quality of data fed into them during the training process. With the main objective of enabling experimentation, we have:

- Collected datasets for the spoken languages of the project
- Collected parallel data of oral languages and sign languages (e.g. text and glosses and videos/images)
- Collected lexicons for oral and sign languages (e.g. glosses, sign dictionaries and so on)

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

What types and formats of data will the project generate/collect?

In relation to UPF-GTI tasks

- The most common data format to exchange *animations* is BVH, and UPF-GTI has been using it.
- *Software* developed by UPF-GTI will be web-based and graphics oriented, meaning heavy use of Javascript and WebGL.
- BML *specification* extension will be an XML Schema; SiGML is a machine readable form of HamNoSys notation

TALN-UPF

The target datasets have different formats (raw, csv, ELAN, etc) and modalities (text, videos, pose information, etc). Thus, these data have been processed[26] to generate formats ready to be fed into neural network models. Both source and processed data follow txt, csv and xml formats in the case of text; and MP4 and jpg/png in the case of videos and images.

Will you re-use any existing data and how?

In relation to UPF-GTI tasks

- So far we did not find *animations* of signs that can be reused and are open; if we discover some or some appear we'll reuse them as much as possible. For the ML system, animations of movements available for free (from Mixamo, https://www.mixamo.com) have been used.
- *Software* developed by UPF-GTI will re-use existing libraries, mostly from WebGLStudio, Open Source graphics framework developed by UPF-GTI (https://webglstudio.org/). Currently we are also using the Open Source MediaPipe (https://mediapipe.dev/), and the open source 3D library Three.js (https://threejs.org/). We have been using a lexicon of signs in NGT which are represented in SiGML
- The extension of the BML *specification* is based on BML

[26] Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, Horacio Saggion. Challenges with Sign Language Datasets for Sign Language Recognition and Translation. LREC 2022: 2478-2487

In relation to UPF-TALN tasks:
Existing SL datasets and corpora have been employed when experimenting with neural machine translation models. Additionally, we experimented with text simplification models and, then, simplification datasets were used in this phase. All data are used to train and evaluate our neural models.

What is the origin of the data?

See the answers above in relation to UPF-GTI tasks
Data for the TALN tasks on language processing has several origins, mainly previous projects and web sites.
UPF-TALN sought available data from previous projects (see Deliverable D3.1) , publications in our research area and/or public repositories (such as https://fundacioncnse-dilse.org).

What is the expected size of the data?

The size of animations, software, and specifications are very small (compared, for instance) with videos. The models of the avatars are larger in size. UPF-GTI has changed the approach with respect to avatars, to facilitate personalisation. We have been using an open source of customisable characters of high quality, and have defined a process to be able to use them within the SignON systems, in particular our own; we have modified our software to include a configuration file related to the avatar in use. The actual personalisation with users is not expected to take place during the course of the project, unfortunately. The avatars will be specifically adapted for mobile use.
UPF-TALN
The size of the text data collected varies from several megabytes to some gigabytes depending on the number of samples in the datasets (see Deliverable D3.1). In the case of video data for SLs, the size of data can range from few to hundreds of gigabytes depending on video quality and number of samples in the dataset.

To whom might it be useful ('data utility')?

UPF-GTI expects that specification, animations and software will be useful for researchers and developers.

The data collected/transformed by UPF-TALN will be shared with researchers for replicability taking into consideration the copyright license for each case.

FAIR data

1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

Data will be stored alongside the software in the repositories recommended by SignON (INT CLARIN infrastructure, Github). The mechanisms and conventions of those repositories will be used. Metadata (e.g. keywords) will be added to identify the project and type of content.

What naming conventions do you follow?

Meaningful names will be adopted for identification of resources.

Will search keywords be provided that optimize possibilities for re-use?

Yes. Keywords will be used for the sake of organization and fairness.

Do you provide clear version numbers?

Versioning will be based on the specification of the repository.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

The first goal of UPF-GTI is that their outputs have been tested for quality before any public release. We expect that the animations, software, and specification extension mentioned earlier will be tested and improved during 2022 and the first part of 2023, and will be only internal. In this testing process we will be exploring the different issues mentioned above; we expect metadata, version numbers, etc. being available where outputs will be located, which is not completely decided yet, with URIs at UPF services linking to the specific locations of outputs (such as GitHub probably in the case of software). By October 2023, the testing is almost complete, and the tools will be released by early November 2023.

In the case of UPF-TALN, we collected metadata in an automatic way, including videos and text. Due to the high expenses in data annotation, automatic techniques were employed to annotate extra information such as body keypoints in videos and linguistic features for text. Additionally, we manually annotated linguistic features for gloss text.

2. Making data openly accessible

Which data produced and/or used in the project will 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.

Data from external sources and enriched for the purpose of training algorithms will be made available following the constraints and restrictions (licences) of the original. Models derived from or trained on external sources will be made available only if the sources on which the algorithms were trained allow us to do so.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

The project has already set-up repositories for sharing software and data (GitHub, INT CLARIN infrastructure) and we intend to use them.

What methods or software tools are needed to access the data?

In case specific methods are needed for reading/transforming the data, they will be made available.

Is documentation about the software needed to access the data included?

README files will be provided and documentation on the open software will be made available.

Is it possible to include the relevant software (e.g. in open source code)?

Yes.

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.

The specifications of the repository concerning metadata specification and storage will be followed.

Have you explored appropriate arrangements with the identified repository?

The project has already set-up the repositories.

If there are restrictions on use, how will access be provided?

Login is required.

Is there a need for a data access committee?

At this time no, but this will be revised periodically.

Are there well described conditions for access (i.e. a machine readable license)?

Links to adopted licences will be provided.

How will the identity of the person accessing the data be ascertained?

UPF-GTI intends to make the outputs mentioned (animations, software, and specification extension) openly available.

3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

Some of the data collected and produced during our research might be of interest to the research community and, therefore, shared. In these cases, we follow standards in our respective domains of expertise to encode text, images and other formats. UPF-GTI plans to share via GitHub as we have been doing in recent times.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

In case a vocabulary is provided by the repository it will be used. Additionally free keywords will be employed to maximize discovery.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

See above.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

UPF-GTI will be using standards such as those of the web, BVH, the WebGL API, etc. to make our outputs as interoperable as possible; we'll be paying attention to the evolution of the standards to ensure that interoperability is as future-proof as possible. Support for some additional exchange formats has been included in the software. In the work for NMFs, the work will be based on standard Action Units (FACS). Ensuring that SiGML is supported has been included as a result of the project. Specifications of interoperable skeletons and facial blendshapes to support better virtual signing will be openly released and documented by the end of the project.

4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

We shall study the most appropriate licences to allow re-use.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be made available to project partners as soon as acquired. Data used for experiments and publication of scientific articles will be made available as soon as the articles are published in case there are no restrictions for using it.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Some of the data may be restricted in case exploitation is sought. Generally, the aim is to share the data to allow reproducibility and progress in the field.

How long is it intended that the data remains re-usable?

It will be decided later on.

Are data quality assurance processes described?

In case of human annotations, data quality will be monitored by measuring agreement between annotators.

Each data collection effort will publish its own data quality assurance processes.

UPF-GTI will be making their outputs open, and available through widely used means such as GitHub in the case of software to facilitate re-use as much as possible. We plan to publish papers to disseminate results as much as possible. We'll be making the avatar newly created publicly accessible. We have been creating specifications to facilitate interoperability and re-use, and will make it public and well documented.

Allocation of resources

What are the costs for making data FAIR in your project?

Costs are contemplated in dissemination activities Task 6.1 (WP6) and tasks associated with task 3.1 (WP3).

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

The PI.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

UPF-GTI follows a minimal costs policy, by using widely available platforms such as GitHub and the resources facilitated by UPF. A major cost to make data available for effective re-use is the quality of the data, its testing, its documentation, etc. We will be making sure that all these tasks with major costs are being carried within and covered by SignON, and we'll be looking for funding opportunities to sustain the projects, as we've been able to do in the last 20 years or so.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The data stored at our institution's servers is secured and backed-up on a regular basis. The repositories adopted by SignOn follow strict security procedures including access via password protected login.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes. The repositories adopted by SignON Github and INT CLARIN infrastructure.

UPF-GTI uses the standard security procedures that are provided by UPF, for instance for internal storage of animations, or those of platforms we use such as GitHub. Software and data we manage are not especially sensitive to require additional specific security measures.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

All models derived from collected data can exhibit societal biases present in the collected data. In particular, in the case of language processing technology, text generation systems could generate text which could be seen as inappropriate in specific contexts.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Animations collected by UPF-GTI have a level of abstraction that makes them non-personal data.
Crowd-sourced collection of animations of signs intended (when software is fully developed and tested, likely by mid 2023) should have ethical clearance before carrying it out (including informed consent and other ethical aspects); this needs to (and will) be discussed with the "user partners" of SignON; UPF-GTI will seek this clearance together with them when this process is mature enough and successful. As of

October 2023, the maturity of the tools currently only allows for internal testing, and thus no ethical clearance has been necessary.For text-based experiments to collect information from users, consent forms will be prepared and submitted for approval to the SignON Research Ethics Committee (SignON_D9.1_Ethical Guidelines and Protocols) and the University Ethics committee.

Sign Language Translation Mobile Application and Open Communications Framework

TU Dublin Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
<V1.0>	<irene Murtagh>	<16/11/2021>	<…..>
	<irene Murtagh>	<14/02/2023>	Ethical Aspects and also flag on Clarins Centre
	<irene Murtagh>	<01/12/2023>	

Version #	Implemented by	Revision Date	Description of changes
<V1.0>	<irene Murtagh>	<16/11/2021>	<…..>
	<irene Murtagh>	<14/02/2023>	Ethical Aspects and also flag on Clarins Centre

Table of Contents

Project partners can use the DMP tools and templates provided by their own organisations as long as the guidelines outlined in this report are followed. They may also use the template that is provided by the EC[1] *in the Annex section and that is included in this Appendix.*

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data collection is focused on gathering datasets that will inform the sign language (SL) machine translation (MT) part of the project. We carry out a linguistic analysis of the five sign languages and then leverage this to inform the development of computational models for the MT engine for SLs. This will in turn enable us to generate the various sign languages taking into account the modality difference between signed and spoken languages.

What types and formats of data will the project generate/collect?

We have created an XML description of the Sign_A framework. We are currently expanding out the Sign_A framework specification, which together with Role and Reference Grammar (RRG) will allow us to create a logical structure that will generate synthetic sign, using an extended version of Behavioural Markup Language (BML) to drive an avatar.

Will you re-use any existing data and how?

We will re-use a number of publicly available corpora, identified by WP 3 Task 3.1. We will also be re-using data from existing data corpora already available to the SignON consortium. We will not collect new data.

What is the origin of the data?

The data consists of the data shared within the SignON partners in compliance with GDPR Art. 14. The data is gathered as part of WP 3.2 Task 3.1

What is the expected size of the data?

The XML framework is approximately 50MB.

To whom might it be useful ('data utility')?

The data created by TU Dublin, including the XML description and the linguistic analysis output, will be useful not only for members of the SignON consortium but also for other researchers from other projects working on automatic SLT or on NLP targeting signed and spoken languages.

FAIR data

1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

All data that can legally be shared with the community, after approval by the Ethics Committees responsible, will be made available on Zenodo.org linked to the associated publications, tools, and software, or in a CLARIN data center. Any Open Source Software produced within the project will also be made available on GitHub. Data, tools, and software will all be documented and all will have a DOI assigned by the hosting platforms.

What naming conventions do you follow?

The naming convention will reflect the contents of the respective research data and the year of publication. The naming conventions will be descriptive and align across all project partners for consistency.

Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be provided and will include the name of the project (SignOn) as well as keywords relatable to the project's subject matter. This will be agreed across the partners to increase and standardise findability.

Do you provide clear version numbers?

Zenodo.org and CLARIN data centers enforce a clear versioning scheme, and this will be used for the versioning of data and tools.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Zenodo.org and CLARIN data centers have provisions for assigning metadata. In our metadata schemes, we will make clear the sign translation architecture that we are using.

2. Making data openly accessible

Which data produced and/or used in the project will 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.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

Due to a neural approach to the sign language processing the outputs from TU Dublin are in report format.

How will the data be made accessible (e.g. by deposition in a repository)?

All data and reporting by TU Dublin will be publicly available.

What methods or software tools are needed to access the data?

All data reporting by TU Dublin will be publicly available in PDF format and XML.

Is documentation about the software needed to access the data included?

README files will be provided and documentation on the open software will be made available where necessary.

Is it possible to include the relevant software (e.g. in open source code)?

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.

Specifications relating to the SignON repository, in relation to storage will be followed.

Have you explored appropriate arrangements with the identified repository?

The SignON project has already set-up the repositories available at (https://github.com/signon-project and (https://centres.clarin.eu/centre/22).

If there are restrictions on use, how will access be provided?

During the project and limited to SignON project partners a Data Transfer Agreement is established to exchange such data.

Is there a need for a data access committee?

No.

Are there well described conditions for access (i.e. a machine readable license)?

Yes, Zenodo.org provides well-described conditions for access [2]*.*

How will the identity of the person accessing the data be ascertained?

3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All data will be published in the formats commonly used in the research communities concerned. If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are open access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no open access options, this data format too will be made available alongside the open data format.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that will be hosted in a CLARIN data center, the format requirements of the CLARIN data center will be followed to make data interoperable[3]*.*

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

If available, standard vocabularies for all data types will be used whenever possible to ensure inter-disciplinary interoperability and re-use.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

All data, text, and software will be published under Creative Commons, unless there are contractual or legal reasons that make this not possible.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be made available after the main publication based on the data having been published, or earlier, if possible, but not longer than 6 months after completion and publication of the data. If an embargo is sought beyond these times, the reasons and duration for the embargo will be given.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions, or is prohibited by laws or regulations in the countries of the participants.

How long is it intended that the data remains re-usable?

Data available on Zenodo.org and CLARIN data centers will remain available without a time limit.

To share data after the lifetime of the project, consent of the data subjects will be sought by asking permission for re-use of the data for the following purposes:

- *Linguistic studies about the properties of sign languages, spoken languages, or written texts*
- *Train and test artificial intelligence systems for translation, i.e. machine translation*
- *Train and test artificial intelligence systems for language recognition or synthesis (audio modality)*

This will be included in the informed consent forms.

Are data quality assurance processes described?

Yes, the data quality assurance process is described. The quality assurance processes will include the provision of results along with the data and the peer-review of publications based on the data.

Each data collection effort will publish its own data quality assurance processes.

Allocation of resources

What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting the *data generated by the project. The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo as a repository.*

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

As such, there are no extra costs for making the data FAIR. The costs, in time and effort, to upload data and publications to Zenodo.org or CLARIN Data centers are marginal and covered by the project and its overhead provisions.

Who will be responsible for data management in your project?

The TU Dublin PI will be responsible for data management, including making data and publications FAIR.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

At the project level, SignON has made provisions for long-term secure storage of data and publications in the form of repositories. Data that is uploaded to Zenodo.org and CLARIN data centers will be available without a time limit.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The data stored at our institution's servers is secured and backed-up on a regular basis. The repositories adopted by SignON follow strict security procedures including access via password protected login.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes. Within the repositories adopted by the SignON project.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are legal and ethical limitations on the access of personal data. All other data will be shared under Creative Commons or open-source licenses. The consortium takes the position that special data protection provisions above those required for personal data are not needed in this project.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes. SignON_D9.1_Ethical Guidelines and Protocols set out guidelines and protocols for dealing with this.

For the use case recordings in the hospitality domain that were carried out within the project, participants will be asked for consent to publish the data as a corpus for re-use outside the project after the project end.
The ethics application for the use case recordings is filed at the REC of project coordinator DCU.

For the publication of the data:
For the use case recordings in the hospitality domain that were carried out within the project, participants will be asked for consent to publish the data as a corpus for re-use outside the project after the project end.

Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case.

The corpus will provisionally be made available through INT being a CLARIN data center,and a SignON project partner.

---

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

[2] *See http://about.zenodo.org/policies/*

[3] *see https://www.clarin.eu/content/interoperability*

Sign Language Translation Mobile Application and Open Communications Framework

TCD Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
<V1.0>	Lorraine LEESON	10/11/2022	<…..>
<V1.1>	Lorraine LEESON	23/10/2023	Minor updates to dates associated with when ISL glossaries will be publicly available from Houses of Oireachtas and Justisigns 2 Project Glossary on Gender Based Violence.

Table of Contents

Project partners can use the DMP tools and templates provided by their own organisations as long as the guidelines outlined in this report are followed. They may also use the template that is provided by the EC[27] in the Annex section and that is included in this Appendix.

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

At Trinity, the SignON project team will collect data via focus groups to identify the perceptions and views of deaf, hard of hearing, and hearing people, vis-a-vis use of machine translation for sign languages (MTSL), and with regards to ethical considerations around the use of MTSL with deaf communities.

What types and formats of data will the project generate/collect?

We are collecting data in focus groups. Any video/audio recordings are used only as a stepping stone towards anonymised English language transcripts that we will then use for coding/thematic analysis (Ethics Approval HT66, June 2021- Lorraine Leeson is lead).

We are also collecting some video data that will, with permission, form an open data set. This process has received ethical approval from the Research Ethics Committee at the School of Linguistics, Speech and Communication Sciences, Trinity College Dublin (Ethics Approval HT47, April 2022 - Rachel Moiselle is lead).

For use case recordings in the hospitality domain (Ethics approval request from DCU, January 2023) carried out within the project, participants will be asked for consent to publish the data as

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

a corpus for re-use outside the project after the project end. Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case. The corpus will be made available through a CLARIN data center, provisionally INT being a SignON project partner.

Will you re-use any existing data and how?

We will re-use Irish Sign Language data from projects that we have been engaged in to date, where such permissions for re-use are confirmed.

What is the origin of the data?

- Justisigns 2 Gender Based Violence Glossary (draft content on Centre for Deaf Studies YouTube site). Final data due late 2023 . Source: Erasmus+ Justisigns 2 Project. See: https://justisigns2.com
- Houses of the Oireachtas political language glossary . Final data due 2023 will be available on the Houses of the Oireachtas website.
- Irish Sign Language Linguistics introductions - available on the Centre for Deaf Studies YouTube site and on the Irish Deaf Society's webpage. See: https://www.irishdeafsociety.ie/irish-sign-language/basicintroductionisl/
- D-Signs Project Irish Sign Language signs/short videos developed to support language learners - Some content is on the Centre for Deaf Studies YouTube site (currently undergoing checks. Due 2022). https://www.youtube.com/channel/UCVaVfZvPa16NWvjaupUmHeA
- Hidden Histories project videos. (These are unannotated and have yet to be loaded on the Centre for Deaf Studies YouTube channel)
- Signall Projects (2, III), Medisigns Project, Justisigns 1 and 2 projects interviews/ educational content delivered in Irish Sign Language - with permission of Mr Haaris Sheikh, Project Chair for all named projects. (These are unannotated and have yet to be loaded on the Centre for Deaf Studies YouTube channel).
- Signs of Ireland (SOI) corpus of Irish SIgn Language (Centre for Deaf Studies). Data set originally collected in 2004-5.

What is the expected size of the data?

Circa 1TB

To whom might it be useful ('data utility')?

This data will be of interest to sign linguists, to interpreters, to those seeking source data sets in sign languages. However, the utility of much of the data will be restricted as it remains unannotated.

FAIR data

1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

None of this exists for the pre-existing datasets we are drawing on. We will explore how we can add these to support future use. For new data collected that will be intentionally open, we will ensure that we mark up the data following SignON data expert advice.

What naming conventions do you follow?

The naming convention will reflect the contents of the respective research data and the year of publication.

Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be provided and will include the name of the source project and SignON as well as keywords relatable to the project’s subject matter.

Do you provide clear version numbers?

Zenodo.org and CLARIN data centers enforce a clear versioning scheme, and this will be used for the versioning of data and tools.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Zenodo.org and CLARIN data centers have provisions for assigning metadata, and we will adopt this and adapt if this proves necessary.

2. Making data openly accessible

Which data produced and/or used in the project will 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.

Irish Sign Language data that is publicly available, stemming from other projects that the Centre for Deaf Studies has been involved in, and which has permissions from the relevant project chairs, can be made open. This includes:

- Justisigns 2 Gender Based Violence Glossary (draft content on Centre for Deaf Studies YouTube site). Final data due 2023.
- Houses of the Oireachtas political language glossary. Final data due 2023.
- Irish Sign Language Linguistics introductions - available on the Centre for Deaf Studies YouTube site and on the Irish Deaf Society's webpage.
- D-Signs Project Irish Sign Language signs/short videos developed to support language learners - On Centre for Deaf Studies YouTube site .
- Hidden Histories project videos. (These are unannotated and have yet to be loaded on the Centre for Deaf Studies YouTube channel)
- Signall Projects (2, III), Medisigns Project, Justisigns 1 and 2 projects interviews/ educational content delivered in Irish Sign Language - with permission of Mr Haaris Sheikh, Project Chair for all named projects. (These are unannotated and have yet to be loaded on the Centre for Deaf Studies YouTube channel).

Data that requires additional permissions for public, open sharing, but which can be shared under restrictions for teaching and learning/ research purposes includes:

- Signs of Ireland corpus. When originally collected in 2004-5, it was not possible to stream videos on the internet as we can today, so we did not think to ask for permission to share in this context. Instead, we have permission to use content for teaching and learning purposes, and to share with researchers, and to use in publications/presentations. We will seek to request additional permissions from participants in 2022 to make some/all of their SOI content open, but until then, we can share on a restricted basis only.

Data that won't be openly shared will be data collected from focus groups associated with Ethics Approval TT66 (Leeson). All details will be translated to English and anonymised. Anonymised texts can be shared with project partners as per the terms of Research Ethics approval secured from the School of Linguistics, Speech and Communication Sciences Research Ethics Committee, Trinity College Dublin, 2021. To access this content, a form needs to be completed by each person accessing the data. The contact for permissions is Prof. Lorraine Leeson (leesonl@tcd.ie). We also plan to deposit content in a repository, INT as CLARIN B Centre (https://centres.clarin.eu/centre/22) . We need to spend some time exploring the requirements associated with doing this, and completing relevant metadata to facilitate this.

For focus groups associated with REC approval HT47 (Moiselle), Our intention is to archive focus group videos where participants grant permission for same. Initially the recordings will be housed on the Centre for Deaf Studies YouTube channel. We seek to do this because (1) Horizon 2020 beneficiaries including the SignON consortium are encouraged to make their research data findable, accessible, interoperable and reusable (FAIR) and to follow the principle of data being 'as open as possible, as closed as necessary'. It is in line with this ethos that we intend-with the express (written) permission of all participants-to have these focus groups be an open dataset; (2)The SignON project is adopting a user-centred, community-driven research and development approach. The consortium's methodology is based on a constant exchange of information and ideas between Deaf/hard of hearing communities and technical experts. In archiving the focus groups, we are following best practice in honouring these co-construction principles: our focus groups will contribute to the conversation surrounding an increasingly salient topic in Deaf communities at an important juncture in research in this arena, and publishing the footage online will ensure that the participant's individual involvement in this contribution will be acknowledged accordingly, as opposed to being relegated to a generic, anonymised reference. Further (3), ISL is an under-resourced and under-researched minority language, particularly in terms of digital content. Our focus groups will be capturing conversations about ISL terminology in ISL. Most linguistic research into ISL is published in English: archiving these focus groups will serve as a mechanism towards the process of repatriating the language to the community of origin in this area of research. The archival protocols for the video recordings of the focus groups are clearly outlined in the Participation Information Leaflets and Consent Forms. To preserve the confidentiality of prospective participants, they will be asked for express (written) permission for their identity to be collected and published. Prospective participants will also be asked whether they consent to having their metadata associated with the data files, as well as being acknowledged by name for their contribution to the research.

3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All data will be published in the formats commonly used in the research communities concerned, though we note that there are different approaches that have been taken to the annotation of sign language datasets across different languages communities, across different research groups, and informed by different understandings of sign language structure at different points in time. Where new data is annotated/marked up, and If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are open access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no open access options, this data format too will be made available alongside the open data format.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that will be hosted in a CLARIN data centre, the format requirements of the CLARIN data centre will be followed to make data interoperable[28].

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Standard vocabularies for all data types will be used whenever possible to ensure inter-disciplinary interoperability and re-use, if available.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

All data and software will be published under Creative Commons, unless there are contractual or legal reasons that make this not possible.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be made available after the main publication based on the data having been published, or earlier, if possible, but not longer than 6 months after completion and publication of the data. If an embargo is sought beyond these times, the reasons and duration for the embargo will be given.

[28] see https://www.clarin.eu/content/interoperability

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions, or is prohibited by laws or regulations in the countries of the participants.

How long is it intended that the data remains re-usable?

Data available on Zenodo.org and CLARIN data centres will remain available without a time limit.

To share data after the lifetime of the project, consent of the data subjects will be sought by asking permission for re-use of the data for the following purposes:

- Linguistic studies about the properties of sign languages, spoken languages, or written texts
- Train and test artificial intelligence systems for translation, i.e. machine translation
- Train and test artificial intelligence systems for language recognition or synthesis (audio modality)

This will be included in the informed consent forms.

Are data quality assurance processes described?

Yes, the data quality assurance process is described. The quality assurance processes will include the provision of results along with the data and the peer-review of publications based on the data.

Each data collection effort will publish its own data quality assurance processes.

Allocation of resources

What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting the data generated by the project. Additional work on annotating the Irish Sign Language datasets that are currently unannotated/marked up in any way will be done only insofar as local resources can facilitate vis-a-vis our commitment to the SignON Project. The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Work relevant to making core data that supports our work on SignON will be covered by the SignON project budget. However, we note that there will remain a large body of data which we will not have the resources to mark up in any detail over the life of SignON as this is incredibly time consuming and costly. The costs, in time and effort, to upload data and publications to Zenodo.org or CLARIN Data centers are marginal and covered by the project and its overhead provisions.

Who will be responsible for data management in your project?

The CDS Dublin PI (Professor Lorraine Leeson - leesonl@tcd.ie)) will be responsible for data management, and will work with the CDS team to make data and publications FAIR.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

At the project level, SignON has made provisions for long-term secure storage of data and publications in the form of repositories. Data that is uploaded to Zenodo.org and CLARIN data centers will be available without a time limit.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The data stored at our institution's servers is secured and backed-up on a regular basis. The repositories adopted by SignON follow strict security procedures including access via password protected login.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes. Within the repositories adopted by the SignON project.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are legal and ethical limitations on the access of personal data. All other data will be shared under Creative Commons or open-source licences. The consortium takes the position that special data protection provisions above those required for personal data are not needed in this project.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes. SignON_D9.1_Ethical Guidelines and Protocols set out guidelines and protocols for dealing with this.

Sign Language Translation Mobile Application and Open Communications Framework

VRT Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
1.0	Karim Dahdah	06/12/2021	Initial documentation
2.0	Karim Dahdah	10/12/2023	Final documentation

Table of Contents

Data Summary

VRT has provided a dataset consisting of audiovisual recordings of interpreters using sign language, subtitles to those recordings and, where relevant, the audiovisual recordings to which sign language is provided. These recordings were made in the context of programmes or other productions for, by, or co-produced with VRT. VRT has provided existing recordings of its news broadcasts (both recent and of older date). The dataset also provides a different subset coming from the child news broadcasts (called "Karrewiet") as the sign language being used there is of better quality, as proposed by the local sign language community and project partner VGTC. Videos with children have been adjusted as much as possible (example: blurring faces).

The signed video content has been captured by a Grass Valley LDX compact camera and mixed with the original (news) content via a Grass Valley Kayak mixed with EBox. The video content then has been compressed with the MP4 H.264/MXF AVCi100 codec and has a resolution of 720p or 1080i with a framerate of 25 frames per second. The compressed storage holds around 15 Gb per 10 minute video at a 70% compression ratio.

The subtitles are an interpretation of the sign language shown in the related videos. This dataset has been delivered as XIF files which will hold around 500KB for 10 minutes of the sign language interpreted videos.

The dataset has been used by the technical partners in the consortium to build corpora which has been used to train the machine learning algorithms. The videos are used to extract the different facial and hand expressions. The subtitles have been matched as best as possible according to the sign language shown in the videos. This combination helps to create machine learning algorithms that can understand the extracted facial and hand expressions. Also, the language model, together with the video recordings of the sign language interpreters, makes it possible to create an interactive 3D model. But subtitles also represent some challenges for the training of the language models; they are using grammatically correct Dutch for easy reading and don't always follow the word by word representation of the sign or spoken language.

The datasets themselves are not eligible to be integrated in any software or products developed by partners in the context of the project. Further processing or re-use beyond these purposes is excluded without previous consultation with VRT.

The dataset has a size of around 800 GB and is made available to the consortium partners through the file servers provided by INT. The dataset has been delivered by disk which then have been transferred securely to those file servers.

FAIR data

1. Making data findable, including provisions for metadata

The dataset contains a set of programmes which are produced by, for, or in co-production with VRT. The programmes are divided in folders, each named after the name of the programme for easy referencing. Each folder holds episodes, which are represented by a MP4 file and a corresponding metadata and subtitle file in the standard broadcast XML based format XIF.

The video files use the H.264/AVC or H.264/MXF AVCi100 codec and have a resolution of 720p or 1080i with a framerate of 25 frames per second. These codecs are standards and can be processed by lots of video editor software applications, either via gui based software or by command line applications such as ‘ffmpeg’. To lower the required disk space, a 70% compression ratio has been adapted, providing a good quality that can be used for training purposes.

The metadata provided differs sometimes as it has been extracted directly from VRT’s archive system. Some episodes’ metadata contain names of (public) people, other can have a description of the content of the episode. The metadata is stored as key-value pairs where the key is a human-readable word describing the content.

2. Making data openly accessible

The datasets provided by VRT cannot be made openly available and can only be used in a strictly scientific (non-commercial) context. VRT does not have the necessary clearances to allow for other use of the data than foreseen for the SignON project, because some of the data in the dataset is created for VRT or in co-production with VRT. New clearances have to be made when new projects will be created.

However, after the SignON project, the data will be stored on the servers of INT, for which the organisation will act as the liaison between the scientific community and VRT and will discuss a case by case request with VRT before giving access to the applicant.

3. Making data interoperable

The dataset provided to the SignON project partners are delivered as industry standard files like using the H.264/AVC codec for video files, PNG or JPEG for image files and broadcast industry standard XIF for subtitle files. Metadata is encapsulated in a key-value representation. Most of them follow the Dublin Core ontology standard.

4. Increase data re-use (through clarifying licences)

The data generated from VRT within the project has been made available to the consortium partners for merely scientific purposes and for the project's duration. The data cannot be re-used for other purposes or outside of the project's scope.

VRT's Data Protection Officer and a legal representative will make the decision on whether or not to supply the data to a potential new user case by case. The data sharing agreement with this partner will need to be signed before the transfer of data is initiated. Each request can be sent to VRT's DPO (dpo@vrt.be) or to the SignON project lead at VRT, Karim Dahdah (karim.dahdah@vrt.be).

Allocation of resources

Karim Dahdah is responsible for the data management of provided datasets coming from VRT delivered to the project's consortium partners.

The dataset's use is covered by the Data Transfer Agreement as agreed by the consortium partners and can be used for the duration of the SignON project. Additionally, VRT's DPO and legal department will investigate (on a case by case basis) if and how it can provide an "open dataset" to the project consortium partners after the end of the SignON project or to third party research partners. A consortium partner in need for such dataset has to send a request via e-mail (dpo@vrt.be or karim.dahdah@vrt.be).

Storage for the hosting of the dataset is granted by INT, as this is the shared file server for the SignON project.

Long term preservation and handling is done through the collaboration with INT (see information earlier in this document).

Data security

As a public service it's important for VRT to have transparent processes. This is important for building trust with the media user, to whom the company is creating content for. It explains these processes via public webpages like https://www.vrt.be/nl/info/mijn-privacy/ and https://www.vrt.be/nl/over-de-vrt/nieuws/2019/01/25/hoe-beschermt-de-vrt-jouw-privacy/. This information is provided in Dutch.

The VRT has its own Data Protection Officer in-house who is reachable via the email address dpo@vrt.be.

The raw camera data and post produced audiovisual content is being stored on servers at VRT's premise or in VRT's private cloud. This data is available to employees from VRT through diverse software tools like Adobe Premiere or the online archive system, a commercial software application built by Arvato, to name a few. Access is linked to the employee's personal account. The software tools that help manage media content in-house are being served from the company's in-house servers. The software is being checked regularly on vulnerabilities through the collaboration with ethical hackers. This is explained on the web page https://www.vrt.be/en/responsible-disclosure-policy-english-version/.

The exported datasets that are being used in the SignON project have been made available through a physical disk which has been transferred during a physical meeting between Mr. Vincent Vandeghinste and Karim Dahdah. The access for consortium partners to the datasets made available by VRT are handled by the project's file server, provided by INT, through a standardised process.

Ethical aspects

The dataset that has been created for this project contains personal data from the sign language interpreters, which are part of the recordings. Most of the audiovisual recordings to which sign language is provided, contain personal data from other natural persons that appear in the recordings. VRT respects the applicable provisions set forth in the relevant legislation on data protection, amongst which the General Data Protection Regulation (GDPR).

The sign language interpreters where informed and, where necessary, datasets were adjusted by removing specific audiovisual files if consent of a sign language interpreter was not given. Once the dataset had been created and copied to the file server at INT this request could not be executed anymore. This has been explained thoroughly to the sign language interpreters.

VRT has done its utmost to make children and other vulnerable people invisible, or difficult to analyse by AI systems, in the dataset provided to the project partners. Broadcasts with mainly vulnerable people are not included in the dataset by default.

Sign Language Translation Mobile Application and Open Communications Framework

UGent Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
V1.0	Joni Dambre	<07/02/2023>	<…..>
V2.0	Mathieu De Coster	05/12/2023	Added notes on the used SLR data and remarks on where and how code and trained models will be made available.

Table of Contents

Project partners can use the DMP tools and templates provided by their own organisations as long as the guidelines outlined in this report are followed. They may also use the template that is provided by the EC[29] in the Annex section and that is included in this Appendix.

Data Summary

UGent will not collect or generate data. Publicly available data and/or data collected by other partners in the SignON project will be used to develop and/or evaluate new models for information extraction in Sign Language videos. These in turn will be used as input for Sign Language translation models developed and trained by other partners.

To create Sign Language Recognition (SLR) models, UGent will write code to process the data sets and to train models. The trained models will also be made available so that they can be integrated into the SLR component.

All software will be written as Python code. Trained models will be made available as binary files, more specifically PyTorch checkpoint files.[30]

No novel training data will be collected by UGent. UGent will use publicly available data sets and data that has been collected by other partners in the SignON project.

These data sets are:

- The Flemish sign language corpus (Corpus VGT)[31]
- The sign language corpus of the Netherlands (Corpus NGT)[32]
- The Signs of Ireland corpus[33]
- The British sign language corpus (BSLCP)[34]

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

[30] https://pytorch.org/tutorials/beginner/saving_loading_models.html

[31] Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., Nyffels, H., & Verstraete, S. (2015). Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent ism KU Leuven. www. corpusvgt. be.

[32] Crasborn, O. A., & Zwitserlood, I. E. P. (2008). The Corpus NGT: an online corpus for professionals and laymen.

[33] Leeson, L. (2006). Moving Heads and Moving Hands: Developing a Digital Corpus of Irish Sign Language. *Ireland*, 25-26.

[34] Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., & Cormier, K. (2013). Building the British sign language corpus.

- The LSE SIGN-HUB data[35]

Pre-existing data sets of SignON project partners are shared under the terms of the SignON Consortium Agreement, which all partners have signed up to. Since these data will contain personal data, compliance with GDPR Art.14 will be ensured.

The written software will be limited in size, in the order of magnitude of several MegaBytes (MB). Trained SLR models will be in the order of magnitude of several hundreds MB.

The written software will be useful to anyone who wishes to process sign language data sets and train SLR models. The trained models will be useful to anyone who wishes to use an existing SLR model in a machine learning pipeline.

FAIR data

1. Making data findable, including provisions for metadata

Python code will be made available through the SignON GitHub repository. More specifically, the sign language recognition code will be available via https://github.com/SignON-project (https://github.com/signon-project-wp3/slr-pipeline), the data set processing code will be available at https://github.com/signon-project-wp3/SLR-Dataset-Processing and the SLR component code will be available at https://github.com/signon-project-wp3/slr-component. These URLs will remain persistent and unique.

To distinguish between code written to process the different sign language data sets, we will make use of the commonly known sign language acronyms: VGT for Flemish Sign Language, NGT for Sign Language of the Netherlands, BSL for British Sign Language, ISL for Irish Sign Language and LSE for Spanish Sign Language.

The SLR models will have clear version numbers, following the MAJOR.MINOR.PATCH naming scheme. The SLR models will also be associated with a “commit hash”, allowing interested parties to link the trained model files with the exact state of the code base used to generate them.

[35] The SIGN-HUB project (https://ww3.thesignhub.eu/project), 2020.

2. Making data openly accessible

All relevant code and trained models will be made openly available.

Python code will be made available through the SignON GitHub repository. More specifically, the sign language recognition code will be available at https://github.com/signon-project-wp3/slr-pipeline, the data set processing code will be available at https://github.com/signon-project-wp3/SLR-Dataset-Processing and the SLR component code will be available at https://github.com/signon-project-wp3/slr-component. These URLs will remain persistent and unique.

Trained models will be shared through the IVDNT as CLARIN B Centre (https://centres.clarin.eu/centre/22).

The Python software code can be opened with any text editor, or read through the user's web browser on GitHub directly.

To open the model checkpoint files, Python 3.10 with PyTorch installed is required.

3. Making data interoperable

As part of the SignON project, we have defined a new standard format for isolated SLR data sets. The model checkpoint format is the standard PyTorch format.

4. Increase data re-use (through clarifying licences)

The code will be licensed under the permissive MIT license. It will be available at the end of the SignON project. No restrictions are in place regarding the usability of the code by third parties and the code will remain usable indefinitely.

Allocation of resources

UGent is not collecting or sharing any data in this project (see above).

Data security

UGent is not collecting or sharing any data in this project (see above).

Ethical aspects

UGent is not collecting or sharing any data in this project (see above).

Sign Language Translation Mobile Application and Open Communications Framework

VGTC Data Management Plan 2023

Authors: Caro Brosens

Version #	Implemented by	Revision Date	Description of changes
<V0.1>	Caro Brosens	<03/11/2021>	First version DMP VGTC

<V0.2>	Caro Brosens	<16/11/2022>	Second version DMP VGTC
<V0.3>	Caro Brosens	<18/10/2023>	Final version DMP VGTC

Table of Contents

1. Data Summary

1.1. What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data brought in by the VGTC will contribute to the overall lexical understanding of Flemish Sign Language in the AI. It provides a robust (lexical) basis to train the AI with.

1.2. What types and formats of data will the project generate/collect?

The VGTC has so far brought in several types of data: the content of the online dictionary VGT/Dutch, online webinars, and research reports.

The online dictionary VGT/Dutch provides information on over 10 000 lexemes in VGT, namely a video of the sign, the corresponding gloss, possible translations, a phonological annotation, and the semantic field(s), i.e. domain labels. Important to note here is that only the “confirmed” lexemes will be included, meaning the ones which have been published on the online website, not the entire database.

Besides the dictionary, two research reports translated into VGT were provided, one on the use of classifier handshapes in VGT and the other on how plurality is expressed in VGT. Lastly, the three webinars (co)hosted by the VGTC contain several hours of (semi)spontaneous signing.

1.3. Will you re-use any existing data and how?

All data provided by the VGTC is already existing and publically available in some form or another. The VGTC will not actively be creating new data for the SignON project. However, if during the project VGTC for example hosts a new webinar, this will also be added as soon as possible.

1.4. What is the origin of the data?

The dictionary data comes from years of extensive linguistic and lexicographic research. The webinars and the research reports stem from VGTC’s continuous engagement with the community.

1.5. What is the expected size of the data?

The overall size of the data offered by the VGTC is not enormous. The dictionary is approximately 11 to 12 gb for over 10 000 lexical signs. The webinars and translated research reports are the smallest in size so far, only about 2 gb.

1.6. To whom might it be useful ('data utility')?

The data could be used in a wide variety of applications. It could for instance be used to develop language learning apps, flesh out sign language courses, train AI, ….

Within the project, however, the data will be useful to train algorithms with regards to NLP and MT. More specifically, it will provide the AI with a robust lexical basis to start from.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

What naming conventions do you follow?

- The lexemes are automatically allocated an identification number. The video name = GLOSS+SEQUENCE LETTER+IDENTIFICATION NUMBER. E.G. PARAPLU-A-9001

Will search keywords be provided that optimize possibilities for re-use?

- No.

Do you provide clear version numbers?

- As of yet irrelevant since there are no different versions (yet). When - in the future - updated versions are provided clear version numbers will be allocated.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

- Metadata of the signer: gender, date of birth, place of residence, …
- Spontaneous/scripted signing or translation

2.2. Making data openly accessible

Which data produced and/or used in the project will 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.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

- The dictionary has been added to the CLARIN ('Common Language Resources and Technology Infrastructure') repository through INT. INT will create metadata CMDI files which will be harvested by the CLARIN Virtual Language Observatory. As such, the dataset will become findable for CLARIN users.

What methods or software tools are needed to access the data?

- Only a video player, a word processor and spreadsheet software are needed to access the data.

Is documentation about the software needed to access the data included?

- No.

Is it possible to include the relevant software (e.g. in open source code)?

- No, but not relevant.

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.

- The data will be added to the CLARIN repository through INT. INT is a certified CLARIN-B center for data depositing.

Have you explored appropriate arrangements with the identified repository?

- Yes, all different licensing options were discussed.

If there are restrictions on use, how will access be provided?

- The use of the data is restricted to scientific use, commercial use might be permitted if directly discussed with the VGTC. The CLARIN infrastructure automatically provides access to the data by academic users in the CLARIN member countries.

Is there a need for a data access committee?

- No.

Are there well described conditions for access (i.e. a machine readable license)?

- Yes, this will be handled through INT.

How will the identity of the person accessing the data be ascertained?

- CLARIN uses federated single sign on. Details are described in https://www.clarin.eu/content/federated-identity

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

- No new data will be produced by the VGTC.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

- The terms used to describe certain concepts (i.e. gloss, annotation, ...) are common in the field of sign linguistics and are to be understood in that context.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

- Yes.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

- As this is not relevant, it is not necessary.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

- The data will be accessible for scientific use. Licenses for commercial use can be discussed.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

- The data will be available for re-use immediately. No embargo will be put on the data.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

- For scientific use, yes. Commercial use will remain restricted, as VGTC maintains full rights.

How long is it intended that the data remains re-usable?

- The dictionary data remains available indefinitely. Just as all natural languages live and change, the online VGT dictionary is also subject to change and a living document. Arrangements will be made for regular updates to be made to the version on the CLARIN server.

Are data quality assurance processes described?

- Our own methodology based on extensive linguistic research and experience will serve as data quality control. These documents can be consulted upon request.

Each data collection effort will publish its own data quality assurance processes.

Allocation of resources

What are the costs for making data FAIR in your project?

- Since this regards pre-existing data, the costs are low. Translating the informed consent forms to Flemish Sign Language is one of the biggest tasks at hand but since this is all done in-house it is covered in personnel costs.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

- The costs are covered in our organisation's personnel costs.

Who will be responsible for data management in your project?

- The current linguistic researcher, Caro Brosens.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

- The practical everyday use of VGT within the community dictates the content and the extent of the dictionary. VGTC merely describes the language as it exists in Flanders. As was mentioned before, the dictionary will remain available indefinitely and will be updated when needed as the language used by the community evolves and changes.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

- A back-up is regularly made of the data, both automatically and manually.
- When sensitive data is being transferred it is encrypted. The key to the encryption is imparted through a different medium.

Is the data safely stored in certified repositories for long term preservation and curation?

- The data has now been transferred to SignON, from there it will be added to the CLARIN server. .

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

- Important to note is that (as with any sign language data) the signers are recognisable in the footage, which makes this kind of data more sensitive. This means extra care with regards to privacy needs to be taken. Before transferring any data to the project, all participants in said data were contacted and provided with informed consents in both Dutch and VGT. We provide informed consents in both Dutch and VGT.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

- Yes.

Sign Language Translation Mobile Application and Open Communications Framework

Radboud University Data Management Plan 2023

Authors: Aditya Parikh, Henk van den Heuvel

Version #	Implemented by	Revision Date	Description of changes
V1.0	Henk van den Heuvel	20/10/2021	First version of DMP RU
V2.1	Aditya Parikh	20/10/2022	Additions, Changes
V2.2	Louis ten Bosch	21/10/2022	Additions, Text Edits
V2.3	Henk van den Heuvel	27/10/2022	Additions, Text Edits, Changes
V2.4	Aditya Parikh	28/10/2022	Final Version 2022 of DMP RU
V3.1	Aditya Parikh	26/10/2023	Additions, Text Edits
V3.2	Henk van den Heuvel	27/10/2023	Final Edits
V3.3	Aditya Parikh	27/10/2023	Final Version 2023 of DMP RU

Table of Contents

1. Data Summary

1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of the data collection is to gather resources to be used in research into optimal methods of natural language processing (NLP) and automated speech recognition (ASR) for the development of a smooth communication service that uses MT to translate between verbal languages and facilitates the exchange of information among deaf and hard-of-hearing (DHH) and hearing individuals. That is, the data will be used to build new models, update and evaluate existing models, improve user experience, etc. Models and algorithms that will exploit the data will be designed and developed such that the information contained therein cannot be deduced to any individual.

1.2 What types and formats of data will the project generate/collect?

In the SignON project, video and audio data from the people who are normally hearing and deaf with hearing difficulties participants will be collected and processed.

The audio data from the individuals who can hear normally could be recordings of predefined sentences. These audio data are collected from open-source platforms like Common-voice, Google Fleurs, etc with particular licenses and agreements. The audio data collected from the Common-voice platform (https://commonvoice.mozilla.org/nl) has metadata such as speaker age group, gender, accent, geographical location, etc.

Apart from the audio data, we also collect a large amount of text data for the language models. For the text data, we are dependent on the Open Source data.
The audio data for hard-of-hearing people is being collected with the help of a special-purpose SignON recording application named "SignON ML" developed by project partners Fincons and MAC. In this application, the sentences/prompts are being provided via Email to the participants, and the spoken audio data is recorded and collected after explicit permission from the participant through the app. Each participant is provided with an identifier for internal use in the system, while the personal information from the speakers includes participant's hearing status, gender and age group and spoken language, which are relevant for research purposes.

Participants are recruited via national and local organizations for deaf and hard-of-hearing people. Via these organizations participants receive the information sheet at home with instructions on how to use the SignON ML app. They enter their metadata and give their explicit consent for the recordings in the mobile application. This data is separated from the speech recordings through pseudonymization.

The recordings and the associated metadata are stored on a locally secured server of one of the project partners (IVDNT). The SignON recording app is GDPR compliant in storing personal data (this is part of the requirements of the EC for funding the SignON project). The data will be transported to RU data storage locations on Ponyland (https://ponyland.science.ru.nl/doku.php?id=wiki:ponyland:about) for processing purposes (ASR evaluation) and the Radboud Data Repository (https://data.ru.nl/) using encrypted connections.

Participants of these recordings are asked for consent to share their recordings and the associated transcriptions and metadata with the partners of the SignON project (for which the project has a data transfer agreement for joint controllers in place), and to publish the data as a corpus for re-use outside the project after project end. Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case.

Further we make audio recordings with internal participants i.e. consortium members on 15 sentences selected from the Horeco dataset stored at the SignON CLARIN B center at IVDNT. These will be internally used and exchanged among project partners only.

1.3 Will you re-use any existing data and how?

Yes, at Radboud University we are responsible for automatic speech recognition. Therefore, for training and testing the acoustic models and language models, we will be using the existing audio and text data from the open-source platform with individual licenses and agreements as mentioned in section 1.2. The languages involved in Automatic Speech Recognition are Northern and Southern Dutch, English, Spanish and Irish.

1.4 What is the origin of the data?

The required data for training and testing ASR systems are the raw audio and text data. In the case of audio, it can be a predefined sentence/prompt spoken by a speaker. Speech corpora originated from established data centers such as LDC (www.ldc.upenn.edu) or they can be chosen from open-source platforms. Similarly, the text corpora for language modeling in ASR originates from LDC or open-source platforms, and usually come with a particular license and/or agreement of use. An open-source platform like Common-voice collects speech audio from various speakers/users around the world, speaking different languages via their website/platform. This platform facilitates a speaker to read aloud randomly generated prompts and record the speech audio. The recordings are then verified and curated by other users/speakers. These types of open-source datasets are constantly growing.

Corpora from LDC and open-source platforms usually contain speech from speakers without hearing problems or speech production problems. The audio data from deaf and hard-of-hearing people will be collected with the help of the "SignON ML'' mobile application in which the user will be provided with predefined prompts/paragraphs. They will be provided with detailed instructions to record their speech.

Their speech will be saved in the database as described in section 1.2. Along with the audio, some metadata like speaker age group, their hearing status, gender, etc. will also be stored.

Pre-existing data sets of SignON project partners are shared under the terms of the SignON Consortium Agreement, which all partners have signed up to. Since these data will contain personal data, compliance with GDPR Art.14 will be ensured.

1.5 What is the expected size of the data?

The size of the research data could be 10 GB to a few hundred GBs. The size of the research data that we have planned to collect ourselves via “SignON ML” mobile application appears to be limited and does not exceed 5GB..

1.6 To whom might it be useful ('data utility')?

The data collected will be useful for anyone interested in speech recognition in connection to the DHH community. However, access to the original data provided by SignON partners may be limited due to privacy and IP concerns.

2. FAIR data

2.1 Making data findable, including provisions for metadata

2.1.1 Are the data produced and/or used 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)?

All data that can legally be shared with the community, after approval by the Ethics Committees responsible, will be made available on Zenodo.org linked to the associated publications, tools, and software, or in a CLARIN data center. Any Free Libre Open Source Software (FLOSS, GPLv3) produced within the project will also be made available on GitHub (https://github.com/signon-project). The resulting models (acoustic models, language models) will be shared via SignON’s dataserver at IVDNT. Data, models, tools, and software will all be documented and all will have a DOI assigned by the hosting platforms.

Suppose the original data underlying a publication cannot be made available due to lack of consent. In that case, aggregated (and in this way anonymized) data will be made available (e.g. in the publication). In this, SignON will seek every possible strategy against re-identification.

2.1.2 What naming conventions do you follow?

The naming convention will reflect the contents of the respective research data (catalog number, version number, date of the last update, speech audio, and sign language video) and the year of publication.

2.1.3 Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be provided to optimize possibilities for re-use. Search keywords will include the name of the project (SignON) as well as keywords relatable to the project's subject matter such as ASR, SLR, etc.

2.1.4 Do you provide clear version numbers?

Zenodo.org and CLARIN data centers as well as RDR (Radboud Data Repository) enforce a clear versioning scheme and this scheme will be used for the versioning of data and tools. Also, the open source datasets like Common-voice which are constantly growing come with unique version numbers stating the most recent updates in the dataset.

2.1.5 What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Zenodo.org and CLARIN data centers have provisions for assigning metadata. In our metadata schemes, we will make clear whether the data are from deaf or hearing signers (incl. age of acquisition of sign language) and from hearing people or DHH people. For the audio data, generated from data warehouses as LDC has a README text file that contains general information about the dataset as well as the structure of a dataset. The audio data acquired from the open-source platform like common-voice comes with a tab-separated value (TSV) file with the speaker's age group, accent, geographical location of the speaker, gender, and respective transcript of the audio file.

2.2 Making data openly accessible

2.2.1 Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why separating legal and contractual reasons from voluntary restrictions.

SignON investigates sign language and speech from DHH (Deaf and Hard of hearing) people. Both at the European and national levels, privacy regulations require that researchers secure ethical approval and informed consent before the publication of data from human participants is permitted. As a result, participant data can only be shared if informed consent to share the data was given by the data subjects. This will always be on a restricted access basis for the raw Audio-Video data that we collect for this project at Radboud University.

2.2.2 How will the data be made accessible (e.g. by deposition in a repository)?

All data that can be collected without the approval of the Ethics Committee will be made openly available from Zenodo or CLARIN and indexed in OpenAIRE.

Participants of the audio recordings through the "SignON ML" will be asked for consent to share their recordings and the associated transcriptions and metadata with the partners of the SignON project (for which the project has a data transfer agreement for joint controllers in place), and to publish the data as a corpus for re-use outside the project after project end. Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case.

For the other data, informed consent as approved by the Ethical Committees will be taken for making data accessible. For Audio-Video data this will typically imply restricted access. At Radboud University the Radboud Data Repository will be used for such data.

Data will be of varying natures and published in commonly used, standard formats. All data will be accompanied by documentation of how to read and use it. If necessary, the required software tools will be described or included.

For educational purposes, the project records presentations of workshops and webinars when the presenter agrees to record. These recordings will be made fully public on Zenodo.org after curation and signed approval by the presenters.

2.2.3 What methods or software tools are needed to access the data?

All open data and publications will be stored on Zenodo.org, in CLARIN data centers such as the CLARIN B Centre at IVDNT which are supported by OpenAIRE and H2020. The models will be stored in such formats that commonly used speech recognition frameworks such as Kaldi can integrate them. The models generated by the end-to-end deep learning approach will be stored in such a way that they can be readily called for decoding via a straightforward python script request. Experimental results and open-source models will be shared via HuggingFace or the CLARIN B Centre at IVDNT.

2.2.4 Is documentation about the software needed to access the data included?

No additional documentation is needed to open the research data.

2.2.5 Is it possible to include the relevant software (e.g. in open-source code)?

Yes, it is possible to include the relevant software.

2.2.6 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.

All data and publications will be stored on Zenodo.org, or in CLARIN data centers which are supported by OpenAIRE and H2020. Typically the CLARIN B Centre at project partner IVDNT will be used for sharing the resulting materials.

For the use case recordings in the hospitality domain that were carried out within the project, participants are asked for consent to publish the data as a corpus for re-use outside the project after the project ends. Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case. The corpus will be made available through a CLARIN data center, provisionally IVDNT being a SignON project partner.

Any Free Libre Open Source Software (FLOSS, GPLv3) produced within the project will be made available on GitHub (https://github.com/signon-project). All the code and its documentation will also be stored on GitHub as a private repository. Such repositories for all the work packages have been created by a SignON partner FINCONS.

Models that are developed within the scope of the project will be stored and shared in a common secure space at the data servers of project partner IVDNT which is a CLARIN B centre, which also has provisions for assigning metadata.

2.2.7 Have you explored appropriate arrangements with the identified repository?

The CLARIN B Centre at IVDNT has an assigned deposit location for SignON materials.

2.2.8 If there are restrictions on use, how will access be provided?

At Radboud University, the Radboud Data Repository (RDR) will be used for data that will not be openly accessible.

During the project and limited to SignON project partners a Data Transfer Agreement is established to exchange such data.

2.2.9 Is there a need for a data access committee?

For data that is not openly accessible, if possible, restricted access limitations will be formulated in user licenses DUA (Data User Agreements). Requests to access such data outside the SignON project will go via the research director of CLS at Radboud University and will require acceptance of the DUA.

2.2.10 Are there well-described conditions for access (i.e. a machine-readable license)?

Yes, Zenodo.org provides well-described conditions for access (see http://about.zenodo.org/policies/). The same holds for the Radboud Data Repository (see https://data.ru.nl/doc/help/helppages/best-practices/bp-selecting-dua.html?7) and so does the CLARIN B Centre at IVDNT (https://ivdnt.org/onderzoek-projecten/clarin/#clarin). For not openly accessible data a DUA (Data User Agreement) will apply (see above).

2.2.11 How will the identity of the person accessing the data be ascertained?

Users are required to register to use the repository.

2.3 Making data interoperable

2.3.1 Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, 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)?

All data will be published in the formats commonly used in the research communities concerned. If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are FLOSS access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no FLOSS (Free/Libre and Open-Source Software) access options, this data format too will be made available alongside the open data format.

2.3.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that will be hosted in a CLARIN data center, the format requirements of the CLARIN data center will be followed to make data interoperable. (see https://www.clarin.eu/content/interoperability)

2.3.3 Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

If available, standard vocabularies for all data types will be used whenever possible to ensure inter-disciplinary interoperability and reuse.

2.3.4 In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

2.4 Increase data reuse (through clarifying licenses)

2.4.1 How will the data be licensed to permit the widest reuse possible?

All data, text, and software will be published under Creative Commons or, for software, FLOSS (Free/Libre and Open-Source Software) licenses, unless there are contractual or legal reasons that make this not possible.

For data that is not openly accessible, restrictions mentioned under 2.2.8 and 2.2.9 will apply.

2.4.2 When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be made available after the main publication based on the data having been published, or earlier if possible, but not longer than 6 months after completion and publication of the data. If an embargo is sought beyond these times, the reasons and duration for the embargo will be given.

2.4.3 Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions, or is prohibited by laws or regulations in the countries of the participants.

For the use case recordings in the hospitality domain that were carried out within the project, participants are asked for consent to publish the data as a corpus for re-use outside the project after the project ends. Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case. The corpus will be made available through a CLARIN data center, provisionally IVDNT being a SignON project partner.

2.4.4 How long is it intended that the data remains re-usable?

Data available on Zenodo.org and CLARIN data centers will remain available without a time limit.

In order to share data after the lifetime of the project, consent of the data subjects will be sought by asking permission for the re-use of the data for the following purposes:

- Linguistic studies about the properties of sign languages, spoken languages, or written texts
- Train and test artificial intelligence systems for text-to-text translation, i.e. machine translation
- Train and test artificial intelligence systems for spoken language recognition or synthesis (audio modality)

This will be included in the informed consent forms.

2.4.5 Are data quality assurance processes described?

Yes, the data quality assurance process is described. The quality assurance processes will include the provision of results along with the data and the peer-review of publications based on the data.

3. Allocation of resources

3.1 What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting the data generated by the project. The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo, CLARIN Data Centres, and the RDR (Radboud Data Repository) as repositories.

Some institutions can offset article processing charges (APC) because they participate in negotiated blanket publication agreements with the publishers.

For this, we will resort to: https://www.openaccess.nl/en

3.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

As such, there are no extra costs for making the data FAIR. The costs, in time and effort, to upload data and publications to Zenodo.org or CLARIN Data centers are marginal and covered by the project and its overhead provisions.

Radboud University has rules and principles in place that require their researchers to make their data FAIR.

3.3 Who will be responsible for data management in your project?

The Principal Investigators (PIs) of the project from the different institutions will be responsible for data management, including making data and publications FAIR.

The coordinator will oversee the implementation. All publications and data that can be published will be stored at Zenodo, IVDNT as CLARIN B Centre (for data and models). Data that cannot be published or has other restrictions will be stored at the institutions or on secure remote storage, at the choice of the hosting institution. Only descriptions and contact information of this latter data will be published to make them findable.

3.4 Are the resources for long-term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

At the project level, we have made provisions for long-term secure storage of data and publications in the form of repositories. Data that is uploaded to Zenodo.org and CLARIN data centers will be available without a time limit. Data that is not open will be stored according to the rules of the RDR (Radboud Data Repository). The use of these long-term repositories does not constitute a cost for the SignON project. The minimum period for data retainment is set to 10 years at Radboud University.

4. Data security

4.1 What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Under GDPR, data subjects have the right to withdraw consent, and that consent must be as easy to withdraw as it was to give. We will therefore ensure that consents are properly managed and can be withdrawn at any time. On the other hand, it must be stressed that data subjects cannot request the deletion of the collected data according to GDPR Art. 17 3 (d) (Art. 17 GDPR – Right to erasure ('right to be forgotten') - General Data Protection Regulation). At CLS of Radboud University, data subjects can claim deletion of their data up to two weeks after data collection.

Unpublished data will be stored in RDR (Radboud Data Repository) which has its own data security provisions. The data will be stored for at least 10 years for instance, for reasons of scientific integrity. Published data will be deposited in Zenodo.org or CLARIN data centers for long-term preservation and curation.

During the project, data generated by partners and provided to RU will be stored in RDR and on the Ponyland HPC servers.

Data exchange between project partners will follow guidelines laid out in the templates for the project's Data Transfer Agreements (D7.9) and will use encrypted file transfer facilities such as FileSender (SURFfilesender: send large files securely and encrypted). For data exchange among the partner institutes a separate SFTP server has also been created. A template for Data Transfer Agreements for joint controllers concerning data obtained from third parties such as broadcast companies has also been provided in D7.9.

4.2 Is the data safely stored in certified repositories for long-term preservation and curation?

Zenodo.org and CLARIN data centers as well as RDR (Radboud Data Repository) have adequate provisions for data security.

Apart from that, Data protection support is provided by the DCU Data Protection Officer, Martin Ward, and the DCU Data Protection Coordinator, Joan O'Connell. They can be contacted at data.protection@dcu.ie.

For queries around the protection of personal data with respect to the SignON project, the queries can be raised at SignON-data-protection@adaptcentre.ie

5. Ethical aspects

5.1 Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action (DoA).

There are strong legal and ethical limitations on the access of personal data. All other data will be shared under Creative Commons or FLOSS licenses. If possible, informed consent for the long-term preservation and sharing of the data will be sought from the data subjects. In agreement with GDPR, all participants in any data collection process will have access to information in plain language and/or signed language, as is their preference. The consortium takes the position that special data protection provisions above those required for personal data are not needed in this project.

As discussed above in DMP, we are not collecting any metadata that relates to the health of any individual - our focus is solely on participants' linguistic backgrounds, accent, gender, age group, and their experiences with and attitudes to MT (Machine Translation).

5.2 Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, informed consent for data sharing and long-term preservation is included in questionnaires dealing with personal data. This will also be implemented for the SignON Use Case speech recordings (see section 1.2). For these recordings long term preservation and sharing is part of the consent asked for. The ethics application for the use case recordings is filed at the REC of project coordinator DCU.

New data that will be collected will require a two-stage process of ethics clearance. That is, when we seek to collect data (e.g. from participants in focus groups, or in creating additional data sets to supplement existing corpora), partners will prepare an ethics application for their home institution, or, if they are a non-university partner, an application for research ethics approval will be submitted via the coordinating partner institution, Dublin City University. Before submitting their application to their institutional research ethics committee, the application will be reviewed by the SignON Ethics Committee.

6. Other Issues

6.1 Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

We will follow local national or departmental procedures for data management that apply to their work because most of the original data that is collected or analyzed is regulated by the local Ethics Committees. In our case, this is EACH - Ethics Assessment Committee Humanities (Ethics Assessment Committee Humanities - Faculty of Arts).

This means that data collected from data subjects will be stored and managed by Radboud University. Use of other publicly available resources such as those obtained through LDC (www.ldc.upenn.edu), ELRA (www.elra.info) or Open Source platforms like Common-voice (https://commonvoice.mozilla.org/en) will be managed as per the end-user license agreements.

Sign Language Translation Mobile Application and Open Communications Framework

KU Leuven Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
V1.0	Tim Van de Cruys	20/11/2022	Update of DMP 2021
v2.0	Myriam Vermeerbergen	29/1/2023	Incorporation of some suggested changes from reviewers + differentiation between the different activities / KU Leuven teams
v3.0a	Bram Vanroy	05/10/2023	Small changes for the technical team
v3.0b	Bram Vanroy	27/10/2023	More information about LDC
v3.0	Bram Vanroy	06/12/2023	Final version

Table of Contents

Data Summary

1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of the data generation carried out by KULeuven is research on automatic sign language translation. That is, the data will be used to build new models, update and evaluate existing models. Models and algorithms that will exploit the data will be designed and developed such that the information contained therein cannot be deduced to any individual. The end goal is the development of models that will translate between different European signed and spoken languages to facilitate the exchange of information among deaf and hard-of-hearing (DHH) and hearing individuals.

Linguistic team KU Leuven

The purpose of the data collection carried out by the linguistic team of KU Leuven is to get more insight into how deaf signers use sign language through video-mediated communication. Since there is no similar study so far for Flemish Sign Language, this exploratory research wants to investigate whether and in which way grammatical mechanisms, tendencies and language use in Flemish Sign Language are influenced by the use of online communication (i.e. communication through laptop and smartphone). The results of this research could support the development of a SignON-application that could take into account the possible singularities of online communication in Flemish Sign Language compared to the singularities of face to face communication in Flemish Sign Language that are partly known. The data collected in this study will only serve the linguistic research and will not be used for building new automatic sign language translation models nor for updating or evaluating existing models.

1.2 What types and formats of data will the project generate/collect?

A) Technical team of KU Leuven

Existing English-only datasets in abstract meaning representation (AMR) will be translated to Dutch, Irish and Spanish to train automatic AMR parsers. The project will generate new models for sign language translation with accompanying model files and programming code.

B) Linguistic team of KU Leuven

New research data will be collected by the linguistic team of KU Leuven. These data will be video recordings containing deaf signers signing in different settings (face to face communication, online communication through laptops and online communication through smartphones). These video files will be analysed with ELAN and will only be used for linguistic research. The video recordings and ELAN files will be stored on protected data storage locations of KU Leuven. Access to these data will be restricted to the involved KU Leuven researchers. This will be identical for the limited set of personal data that will be gathered (name - gender - age- regiolect).

1.3 Will you re-use any existing data and how?

A) Technical team of KU Leuven

We will re-use a number of publicly available corpora and data available through project partners for sign language translation. Data from the Linguistic Data Consortium has also been used through the KU Leuven data agreement between KU Leuven and LDC. Some of the data will be manipulated, specifically translated with Google Translate APIs.

B) **Linguistic team of KU Leuven**

For the purposes of the study concerning online communication by deaf Flemish signers, only new data will be collected and analysed.

1.4 What is the origin of the data?

A) Technical team of KU Leuven

Existing data will be used for the creation of novel models. Pre-existing data sets of SignON project partners are shared under the terms of the SignON Consortium Agreement, which all partners have signed up to. Since these data will contain personal data, compliance with GDPR Art.14 will be ensured. Original AMR data was retrieved from the Linguistic Data Consortium (https://catalog.ldc.upenn.edu/LDC2020T02) will also be used.

B) Linguistic team of KU Leuven

For the purposes of the study concerning online communication by deaf Flemish signers, only new data will be collected and analysed.

1.5 What is the expected size of the data?

A) Technical team of KU Leuven

10-100GB

B) Linguistic team of KU Leuven

Video recordings will be collected of no more than 10 participants. The amount and size of recordings per participant can be adapted according to preliminary analyses and findings.

1.6 To whom might it be useful ('data utility')?

A) Technical team of KU Leuven

The data collected will be useful for anyone who is interested in translating between different European signed and spoken languages to facilitate the exchange of information among DHH and hearing individuals. AMR data is bound to push forward the field of multilingual AMR parsing in Dutch, Spanish and Irish. Model weights and programming code will serve for the research community as well as other interested parties who can find a use for multilingual AMR parsing

B) **Linguistic team of KU Leuven**

Access to the novel collected data will be restricted to the researchers involved at KU Leuven due to privacy concerns.

FAIR data

1. FAIR data

2.1 Making data findable, including provisions for metadata

2.1.1 Are the data produced and/or used 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)?

All data that can legally be shared with the community, after approval by the Ethics Committees responsible, will be made available on Zenodo.org linked to the associated publications, tools, and software, or in a CLARIN data center. Any Open Source Software produced within the project will also be made available on Github. Data, tools, and software will all be documented and all will have a DOI or other type of persistent identifier assigned by the hosting platforms. For the translated AMR data we are in the process of publishing the data through LDC, following licensing and publishing agreements of the original AMR dataset. Model weights will also be shared on the Hugging Face platform for ease of access and discoverability.[36]

2.1.2 What naming conventions do you follow?

[36] https://huggingface.co/models

The naming convention will reflect the contents of the respective research data and the year of publication.

2.1.3 Will search keywords be provided that optimize possibilities for re-use?

Search keywords will be provided to optimize possibilities for re-use. Search keywords will include the name of the project (SignON) as well as keywords relatable to the project's subject matter and involved languages.

2.1.4 Do you provide clear version numbers?

Zenodo.org and CLARIN data centers enforce a clear versioning scheme, and this will be used for the versioning of data and tools. LDC allows for "releases", so even though the AMR data is not expected to have other releases, the data repository does allow for it. Model weights on the Hugging Face platform are also version-controlled.

2.1.5 What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Zenodo.org and CLARIN data centers have provisions for assigning metadata. In our metadata schemes, we will make clear which sign translation architecture we are using. In LDC metadata such as ISBN, DOI, Language, License, Citation, Release Date are included.

2.2 Making data openly accessible

2.2.1 Which data produced and/or used in the project will 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.

SignON investigates sign language (SL), speech and text from DHH and hearing people. Both at the European and at a national level, privacy regulations require that researchers secure ethical approval and informed consent before the publication of data from human participants is permitted. As a result, participant data can only be shared if informed consent to share the data was given by the data subjects. The AMR dataset cannot be made fully open because it is a derivative work from the existing AMR 3.0

corpus (https://catalog.ldc.upenn.edu/LDC2020T02). Therefore it is bound by some restrictions, notably that it should be published through LDC for research purposes.

2.2.2 How will the data be made accessible (e.g. by deposition in a repository)?

All data that can be collected without the approval of an Ethics Committee will be made openly available from Zenodo or CLARIN at the INT CLARIN B center[37] and indexed in OpenAIRE. For the other data, informed consent as approved by the Ethical Committee will be taken for making data accessible. The translated AMR dataset will be published through LDC. Programming code will be made available via the SignON github repository.[38]

Data will be of varying natures and published in commonly used, standard formats. All data will be accompanied by documentation of how to read and use it. If necessary, the required software tools will be described or included.

For educational purposes, the project records presentations of workshops and webinars when the presenter agrees to record. These recordings will be made fully public on Zenodo.org after curation and signed approval by the presenters.

Model weights will also be shared on the Hugging Face platform.

2.2.3 What methods or software tools are needed to access the data?

All open data and publications will be stored on Zenodo.org, LDC, or in CLARIN data centers which are supported by OpenAIRE and H2020. The AMR data can be downloaded from the LDC website when it is available and after conforming to licence requirements. The model weights on the Hugging Face platform can be downloaded as-is in its binary format, but for real utility the `transformers` Python library must be used.

2.2.4 Is documentation about the software needed to access the data included?

No additional documentation is needed to open the SignON research data.

2.2.5 Is it possible to include the relevant software (e.g. in open source code)?

[37] https://centres.clarin.eu/centre/22
[38] https://github.com/signon-project

Yes. This is shared on Github.

2.2.6 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.

All data and publications will be stored on Zenodo.org, in CLARIN data centers which are supported by OpenAIRE and H2020, or the Linguistic Data Consortium repository.

All above platforms also have provisions for assigning metadata.

For the use case recordings in the hospitality domain that were carried out within the project, participants will be asked for consent to publish the data as a corpus for re-use outside the project after the project ends. Since the data is not sensitive in terms of content and metadata, the SignON consortium considered that it can meet the EU requirement for publishing data as open in this case. The corpus will be made available through a CLARIN data center, provisionally INT being a SignON project partner.

2.2.7 Have you explored appropriate arrangements with the identified repository?

Yes. INT is a partner within the project and we have close involvement with CLARIN. Contact was established with LDC to investigate how openly the AMR data can be published.

2.2.8 If there are restrictions on use, how will access be provided?

During the project and limited to SignON project partners a Data Transfer Agreement is established to exchange such data.

2.2.9 Is there a need for a data access committee?

No.

2.2.10 Are there well described conditions for access (i.e. a machine readable license)?

Yes, Zenodo.org provides well-described conditions for access and on other platforms (Github, Hugging Face) machine-readable LICENSE files will be included.

2.2.11 How will the identity of the person accessing the data be ascertained?

Not relevant for the open nature of the data. Access to our data on LDC is only possible through an LDC account however.

2.3 Making data interoperable

2.3.1 Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All data will be published in the formats commonly used in the research communities concerned. If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are open access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no open access options, this data format too will be made available alongside the open data format.

2.3.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that will be hosted in a CLARIN data center, the format requirements of the CLARIN data center will be followed to make data interoperable.[39]. In the LDC, we follow the structure template of the LDC with their required metadata fields.

2.3.3 Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

If available, standard vocabularies for all data types will be used whenever possible to ensure inter-disciplinary interoperability and re-use.

2.3.4 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The compatibility of our project-specific ontologies and vocabularies will be guaranteed through appropriate mapping to more commonly used ontologies.

[39] see https://www.clarin.eu/content/interoperability

2.4 Increase data re-use (through clarifying licences)

2.4.1 How will the data be licensed to permit the widest re-use possible?

All data, text, and software will be published under Creative Common or GNU-based licenses unless there are contractual or legal reasons that make this not possible.

For data that is not openly accessible, restrictions mentioned under 2.2.8 and 2.2.9 will apply.

2.4.2 When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be made available after the main publication based on the data having been published, or earlier, if possible, but not longer than 6 months after completion and publication of the data. If an embargo is sought beyond these times, the reasons and duration for the embargo will be given.

2.4.3 Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

SignON strives to make all data usable by third parties from the start, unless sharing the data is not approved via consent forms, by the Ethics Committees of the participating institutions, or is prohibited by laws or regulations in the countries of the participants.

2.4.4 How long is it intended that the data remains re-usable?

Data available on Zenodo.org, CLARIN, Hugging Face, and LDC data centers will remain available without a time limit.

To share data after the lifetime of the project, consent of the data subjects will be sought by asking permission for re-use of the data for the following purposes:

- Linguistic studies about the properties of sign languages, spoken languages, or written texts
- Train and test artificial intelligence systems for text-to-text translation, i.e. machine translation
- Train and test artificial intelligence systems for spoken language recognition or synthesis (audio modality)

This will be included in the informed consent forms.

2.4.5 Are data quality assurance processes described?

Yes, the data quality assurance process is described. The quality assurance processes will include the provision of results along with the data and the peer-review of publications based on the data.

2. Allocation of resources

3.1 What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting the data generated by the project. The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo, LDC or Hugging Face repositories.

3.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

As such, there are no extra costs for making the data FAIR. The costs, in time and effort, to upload data and publications to Zenodo.org or CLARIN Data centers are marginal and covered by the project and its overhead provisions.

3.3 Who will be responsible for data management in your project?

The KU Leuven PI will be responsible for data management, including making data and publications FAIR.

The KU Leuven PI will oversee the implementation. All publications and data that can be published will be stored at Zenodo, Hugging Face, LDC and CLARIN. Data that cannot be published or has other restrictions will be stored at KULeuven.

3.4 Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

At the project level, we have made provisions for long-term secure storage of data and publications in the form of repositories. Data that is uploaded to Zenodo.org, LDC, Hugging Face and CLARIN data centers will be available without a time limit.

3. Data security

4.1 What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Under GDPR, data subjects have the right to withdraw consent, and that consent must be as easy to withdraw as it was to give. We will therefore ensure that consents are properly managed and can be withdrawn at any time. On the other hand, it must be stressed that data subjects cannot request the deletion of the collected data according to GDPR Art. 17 3 (d)[40].

Unpublished data will be stored at KULeuven, which has its own data security provisions. The data will be stored for at least 5 years for instance, for reasons of commercial or scientific integrity. Published data will be deposited in Zenodo.org, LDC and Hugging Face or CLARIN data centers for long-term preservation and curation.

4.2 Is the data safely stored in certified repositories for long term preservation and curation?

Zenodo.org, LDC, Hugging Face and CLARIN data centers have adequate provisions for data security.

4. Ethical aspects

5.1 Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are strong legal and ethical limitations on the access of personal data. All other data will be shared under Creative Commons or open-source licenses. If possible, informed consent for the long-term preservation and sharing of the data will be sought from the data subjects. In agreement with GDPR, all participants in any data collection process will have access to information in plain language in a written

[40] Art. 17 GDPR – Right to erasure ('right to be forgotten') - General Data Protection Regulation

and/or a signed language, as is their preference. The consortium takes the position that special data protection provisions above those required for personal data are not needed in this project.

As discussed above in DMP, we are not collecting any metadata that relates to the health of any individual - our focus is solely on participants' linguistic preferences and their experiences with and attitudes to MT (Machine Translation).

5.2 Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, informed consent for data sharing and long-term preservation is included in questionnaires dealing with personal data.

For the use case recordings in the hospitality domain that were carried out within the project, participants will be asked for consent to publish the data as a corpus for re-use outside the project after the project end. The ethics application for the use case recordings is filed at the REC of project coordinator DCU.

5. Other Issues

6.1 Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No.

Sign Language Translation Mobile Application and Open Communications Framework

EUD Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
V0.1	Davy Van Landuyt	09/10/2023	

Table of Contents

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The EUD's focus groups will, through a grid of questions, contribute to understanding user needs and practices and discuss how the SignON mobile app could be used in everyday life and, by extension, support SignON's research and develop the SignON communication service that uses machine translation to translate between sign and spoken languages.

What types and formats of data will the project generate/collect?

We will collect video data, i.e. video recordings of the focus groups. This video data is not anonymised, and remains private. EUD will transform this raw data by translating the information from International Sign and national sign languages into written English. Once the information has been transcribed, we filter out personal information through the anonymisation process. To anonymise data in transcriptions, participants' names are replaced by codes that indicate the language group that they belong to and an arbitrary number. For example, a participant from a Flemish Sign Language (VGT) focus group may be anonymised as VGT3 in the transcripts. We also censor other possible identifiers e.g. when a participant shares for example the city where they live.

We also collect metadata from participants with respect to age, gender, name and contact details. These metadata are kept separately from video recordings and anonymised transcripts and not shared outside of EUD.

This processed information is being used to feed the deliverables within WP1, as well as academic research output from the SignON project. All data that is shared within the deliverable and among project partners is in written anonymised form.

Will you re-use any existing data and how?

We do not use existing data. Only the data shared during the focus groups is used.

What is the origin of the data?

The data originates from the focus groups of the EUD, which were conducted in the framework of the SignON Project.

What is the expected size of the data?

The total duration of the focus group interviews is 8 hours and 7 minutes. This resulted in a transcription of 49 pages in written English.

To whom might it be useful ('data utility')?

The data collected in the focus groups will benefit the SignON research, and be of immediate use to project partners working to develop the SignON app and service. Outside of this, the data will be of interest to researchers, developers and policymakers working on topics concerning Sign Language Machine Translation, as these data will outline key concerns, opinions and requirements of European deaf communities when it comes to these technologies.

FAIR data

1. Making data findable, including provisions for metadata

The names of the participants have been replaced by the following identifiers: the use of the sign language given by a code: LSE for Spanish Sign Language followed by the number (e.g. LSE1). There is no other naming convention. The anonymised transcriptions won't be publicly findable. Even when we applied anonymisation measures, some of the participants might still be identified by someone that knows the local deaf community well and reads the transcriptions. We can give access to the transcriptions on a case-by-case basis: 'as open as possible, as closed as necessary'.

2. Making data openly accessible

The data (in the form of the anonymised transcriptions) will be shared on Google Drive for the use of the SignON consortium partners but will not be distributed to the general public. However, transcripts in their anonymised form will be shareable to the wider research community upon request and with permission of participants. We will also provide an overall summary of the conducted activities and results on the SignON website (www.signon-project.eu) and share via social media. Academic publications and conference presentations will disseminate the findings to a wider academic audience.

3. Making data interoperable

We plan to keep the video data and the personal data for the life of the SignON project. This raw data will be deleted at the end of the project. Only the anonymised transcripts will be kept within EUD.

4. Increase data re-use (through clarifying licences)

N/A

Allocation of resources

We plan to keep the video data and the personal data for the life of the SignON project. This raw data will be deleted at the end of the project.

Data security

The raw data (video recordings) and metadata (participants age and gender) are stored and saved in files kept secret and controlled only by EUD and DCU. These files are stored in a specific protected cloud folder which requires special access permissions for the EUD staff involved in making and transcribing these recordings. The processed data is stored in Google Drive within the SignON Consortium.

Ethical aspects

We will share an Informed Consent Form in which we will explain what the focus group interview entails and what is being done with all the shared information. For example, we will mention that only the EUD researchers will be able to watch the recordings of the meetings. They will note down all the answers of the participants according to the asked questions. These notes are then compared with the notes taken during the meetings.

All data will then be rendered anonymously and compiled into a report for the project partners. The data is managed and accessed only by EUD's and DCU'S researchers. No one else is permitted to access this data. That is why the use of the data is managed solely by the EUD and DCU researchers from start to finish and the translation work of the video data into text is not outsourced. All participants participate knowingly and have been informed of their freedom to withdraw either before, during or after the focus group and are not forced to participate. The subject matter is not sensitive and will not be such as to impact physically, morally, socially or psychologically. There is no risk for the participants or researchers involved in the focus groups.

There are no payments or incentives to participants for the involvement in the focus groups.

The EUD researchers don’t have any personal interests (nor commercial or financial interests).

Sign Language Translation Mobile Application and Open Communications Framework

TiU Data Management Plan 2023

Authors:

Version #	Implemented by	Revision Date	Description of changes
<V.1.0>	Mirella De Sisto	<24/10/2023>	First version

Table of Contents

Project partners can use the DMP tools and templates provided by their own organisations as long as the guidelines outlined in this report are followed. They may also use the template that is provided by the EC[41] in the Annex section and that is included in this Appendix.

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data collection focuses on gathering resources that can be employed in training and evaluation of machine translation (MT) models. In addition, the data are used for investigating linguistic aspects of sign language use. TiU has been using pre-existing data, either from the data available to the SignON consortium or from external sources, as well as collecting new data for the creation of two parallel multimodal corpora.

The data available to the SignON consortium have been used to compare the annotation conventions of the various corpora and to identify common characteristics and divergence points. The purpose of these tasks was to contribute to the definition of a common format for the corpora used in the project. Thisfacilitates the interoperability of the data and of the models that are be developed with them.

The purpose of the newly created parallel corpora is to provide high quality data for training models as well as gold standard data for evaluation purposes. The advantage of multimodal parallel corpora of both signed and spoken languages is the availability of parallel text and signed videos which well support MT tasks. In addition, one of the newly created corpora also contains parallel data of two signed languages, i.e. Sign Language of the Netherlands and Flemish Sign Language, which facilitates research on signed language to signed language MT and on cross-language comparison.

What types and formats of data will the project generate/collect?

- Collecting: within the project TiU may collect video, audio and text data for the purpose of training MT models. Such models can operate on any sequential data and thus we may develop

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

audio, video or text to audio, video or text translation models. Various formats will be used, depending on the native format of the available data.

- Generating:
  - Through processing and formatting existing data TiU will generate derivatives of the existing resources that are suitable for training and inference with MT systems. These derivatives range between minor formatting (e.g. converting to unicode) to more significant formatting and organisation of the data. However, the original data can be reconstructed by reversing the employed steps.
  - Metadata generated through using data analytics tools.
  - By employing tools and techniques for data modification we may reformat the data into a version that can not be used to reconstruct the original data. For example, text compression, principal component analysis, etc.
  - MT and other data-driven models can generate new data, such as backtranslated data to be used for augmenting existing corpora.
  - For the creation of the two parallel corpora, TiU has been collecting videos in mp4 format, written text, metadata (i.e. item identifier and anonymised signed language translator identifier[42] - for instance, "P2") and EAF files (videos + a tier containing the corresponding Dutch translation).[43]

Will you re-use any existing data and how?

We will be re-using data from existing data corpora already available to the SignON consortium. The data might be implemented or reformatted but without modifying its original content. We will not collect new data.

If new, openly accessible data is made available, we will collect it and use it for developing new and updating existing MT models (including the InterL-E and InterL-S).

What is the origin of the data?

The preexisting data consist of the data shared within the SignON partners in compliance with GDPR Art. 14.

[42] This is done for training purposes.

[43] All videos have been created after receiving ethical clearance from the Research Ethics and Data Management Committee of Tilburg University. Authors of the videos have agreed to their distribution (either CC-BY-NC or CC-BY license) by signing an informed consent form.

The parallel corpora data consist of either newly produced translations into a signed language or written text or of already existing signed language videos.

What is the expected size of the data?

The size of the sign language data sets are limited and vary across the corpora.

To whom might it be useful ('data utility')?

The data will be useful not only for members of the SignON consortium but also for other researchers from other projects working on automatic SLT or on NLP targeting signed languages.

FAIR data

1. Making data findable, including provisions for metadata

Are the data produced and/or used 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)?

Openly available data that has been collected, processed and used within the scope of this project, will be made available. Data will be distributed via IVDNT as CLARIN B Centre . All data and publications will be stored on Zenodo.org, in CLARIN data centres which are supported by OpenAIRE and H2020.

Proprietary data will be used and distributed according to their licenses.

For internal use, all data is stored in a common location (an FTP server) which is accessible (read and write privileges) by all consortium members.

Software tools and scripts used to process or modify the data will be made available (in publicly accessible repositories) and clearly indicated in publications.

What naming conventions do you follow?

No naming conventions have been decided yet.

Will search keywords be provided that optimize possibilities for re-use?

Yes

Do you provide clear version numbers?

Naming and versioning conventions are yet to be decided. The parallel corpus which has been published clearly states the version number (currently v. 2.0).

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

For the already published parallel corpus metadata has been created, consisting in item identifier and anonymised signed language translator identifier. Metadata for other resources will be created. The exact type and format are yet to be defined.

2. Making data openly accessible

Which data produced and/or used in the project will 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.

All pre-existing data which is publicly available in its original format, will be made openly available in the new format.

For educational purposes, the project records presentations of workshops and webinars when the presenter agrees to recording. In case of recordings with no internal use restriction, these will be made available on Zenodo after curation and signed approval by the presenters. For those recordings for which fully open publication is not possible, due to ethical or legal objections, or because the presenter does not consent to such-publication, other solutions will be sought. The presentation of recordings from data subjects in educational settings will depend on consent for this purpose given by the data subjects, and will be included as an option in the consent forms provided to them.

All data and publications will be stored on Zenodo.org, in CLARIN data centres which are supported by OpenAIRE and H2020. Models will be distributed via our GitHub repository.[44]

Based on written agreement with videos' authors, one parallel corpus is available under the CC-BY NC license, while the other parallel corpus will be made available under the CC-BY license.

[44] https://github.com/signon-project-wp4 and https://github.com/SignON-project

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

Data are and will be made available via Zenodo and CLARIN data centres and European Language Grid platform

What methods or software tools are needed to access the data?

Access to the aforementioned services is needed (typically this includes creating an account and agreeing with their terms and conditions). A file sharing client may be required, e.g. FileZilla. Data may be archived, thus, after download or filetransfer, a software for extracting the data is needed, e.g. winzip, winrar, etc.

Is documentation about the software needed to access the data included?

Readme files with instructions will be provided alongside the data.

Is it possible to include the relevant software (e.g. in open source code)?

Software developed by consortium partners or instructions how to install and set up certain software will be provided.

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.

Data and models will be shared through the IVDNT as CLARIN B Centre. Code will be shared through our GitHub repository.

Have you explored appropriate arrangements with the identified repository?

Repositories have already been set up.

If there are restrictions on use, how will access be provided?

Links to adopted licences will be provided.

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

Yes

How will the identity of the person accessing the data be ascertained?

Each of the services we use employs authentication protocols. We will rely on those.

3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use 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)?

All the data produced is and will be interoperable and published in the formats commonly used in MT.

All data is and will be published in the formats commonly used in the research communities concerned. If available, public guidelines for metadata vocabularies, standards, or methodologies will be followed. Standard data formats for which there are FLOSS access options will be used for data. If other formats are necessary, software to access the data will be added to the repository. If the original data format used within the project is proprietary or has no FLOSS access options, this data format too will be made available alongside the open data format.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For data that are or will be hosted in a CLARIN data centre the format requirements of the centre are followed. After defining common metadata and format, clear information about them are made publicly available.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

We aim to use standard vocabularies for all data types present in our data set in order to allow inter-disciplinary interoperability.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

We will use standard ontologies and vocabularies.

4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

For the pre-existing data the licensing will depend on the data owners. New data are made available via CLARIN, European Language Grid, European Language Equality 2, and the Instituut voor de Nederlandse Taal, after receiving consent from participants.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

TiU is not planning to apply any embargo on the data. The data are or will be directly made available or as early as possible.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

The data resulting from an implementation of already publicly available material will be usable by third parties. Most of the pre-existing data is available for scientific use only. In case of new data produced by TiU, these will be usable by third parties. How long is it intended that the data remains re-usable?

The availability of new data produced by TiU at the IDVNT and on the CLARIN platform will ensure long-term availability.

Are data quality assurance processes described?

In case of human annotations, data quality will be monitored by measuring agreement between annotators. For newly translated data, a revision phase of the newly produced material will take place. For automatic annotations, cleaning, formatting and other preprocessing, automatic as well as human evaluation will be used to ensure high data quality.

Each data collection effort will publish its own data quality assurance processes.

A common quality assurance process will be agreed on and followed. If this process cannot be followed due to the circumstances, then yes, the data collection effort will publish the quality assurance process.

Allocation of resources

What are the costs for making data FAIR in your project?

The monetary costs of making data FAIR in the project consist of the publication costs for Open Access publications and the costs of recording, curating, formatting, and hosting of the data generated by the project. TiU can offset article processing charges (APC) because they participate in negotiated blanket publication agreements with some of the publishers.

The remaining Open Access publication costs and the costs of producing and hosting recordings of SignON events (workshops, webinars) are paid out of the dissemination budget. There are no charges for using Zenodo or the IVDNT as repositories.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

From the budget agreed upon in the grant agreement.

Who will be responsible for data management in your project?

Dr. Dimitar Shterionov, the Principal Investigator (PI) of the project from TiU, will be responsible for data management, including making data and publications FAIR.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

TiU has made provisions for long term secure storage of data and publications in the form of repositories. Data that is uploaded to Zenodo and ELG-CLARIN data centres will be available without a time limit. Data that is not open will be stored according to the rules of the owning institution. The use of these long term repositories do not constitute a cost for the SignON project.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The data generated by TiU will be stored in its local secured network facilities, and the specifics of this will be documented. For internal use, data will be copied to the data FTP server of the project so that others can access it. Data within the project will be publicly shared through the IVDNT as CLARIN B Centre[45] by following the Data Transfer Agreements guidelines.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

The collection of videos has received ethical clearance from the Research Ethics and Data Management Committee of Tilburg University. We are not collecting other personal data. All other data will be shared under Creative Commons or FLOSS licenses.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

For new data produced by TiU, informed consent for data sharing and long term preservation has been obtained from the authors of the videos. Original authors of the videos retain the right to have their data removed from the corpora.

[45] https://centres.clarin.eu/centre/22