D9.3 WORKFLOW DOCUMENTS AND MANUALS FOR THE CREATION OF NEW RESOURCES

Work package	WP 9
Task	Task 9.3
Due date	30/09/2022
Submission date	30/09/2022
Deliverable lead	Radboud University
Version	1.0
Authors	Onno Crasborn, Hope E. Morgan
Reviewers	Marc Schulder, Thomas Hanke (UHH), Kearsy Cormier (DCAL)

Abstract	This report contains workflow documents to guide the creation of new sign language resources. Specifically, this is presented via five phases of a language documentation project: (1) planning, (2) data collection, (3) annotation, (4) archiving, and (5) public outreach. Each phase is described in its own manual with targeted recommendations, best practices, and key resources for further information.

Keywords	Sign language documentation, annotated corpora, methods, project management

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

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	19/09/2022	Draft version of D9.3 for reviewer feedback	Onno Crasborn, Hope E. Morgan
V1.0	27/09/2022	First version of document, reflecting reviewer feedback	
V1.0	30/09/2022	Submission	Giacomo Inches (Martel)

DISCLAIMER

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

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

COPYRIGHT NOTICE

© 2022 EASIER Consortium

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

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

DEM: Demonstrator, pilot, prototype, plan designs

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

OTHER: Software, technical diagram, etc.

Funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

This report is intended primarily for those who are planning or currently undertaking a sign language documentation project – specifically, the creation of a corpus (collection) of sign language data on video that will be annotated and archived.

We recommend that readers review this report alongside our previous EASIER report, D9.1 on the '*Definition of minimal contents of dataset for participation'*, in which we discuss quantitative and qualitative benchmarks for language documentation, especially to take advantage of the machine translation tools being built in EASIER.

The current report is a practical guide through the phases of a sign language corpus project. Since so much helpful information has already been written on this topic, our purpose here is provide the main points to consider at each phase of the project, and otherwise to be a helpful reference guide for readers to locate key resources that go into more depth.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE OF CONTENTS

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION

KEY RESOURCES:

- Hochgesang, J. & Fenlon, J., Eds. (2022), Sign language corpora. Gallaudet University Press.

To create sign language resources from scratch is a major effort. Some countries and research institutes have a lot of experience in this domain (Kopf et al. 2021), but many others do not (Morgan and Crasborn 2022). The present documents are aimed at the latter group, in particular European countries that in the future can profit from the newest sign language technologies, such as those developed in this EU project. To make that possible, sufficient language resources must be available.

The aim of this document is to provide insight into the whole workflow of resource creation (focusing on the combination of corpus and lexicon), from planning to exploitation. There is a lot of information available online by now, so we will refrain from repeating detailed information published elsewhere, but provide references. Most of that information is available in the form of open access publications, but there are also book volumes and journal articles available commercially. We will point to key sources in each section below. In particular, we will point to recommendations that emerge from work in EASIER on the harmonization of sign language resources, and best practices emerging from the resource creation efforts on the sign languages that EASIER focuses on.

There are five sections in what follows, representing the five different steps or phases in the corpus creation workflow, from planning to public outreach (Figure 1).

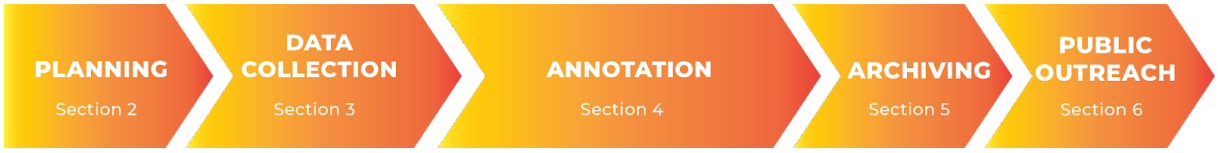

FIGURE 1: STEPS IN CORPUS CREATION WORKFLOW

The focus here is on developing a corpus, but the link to a lexicon is vital, as we will discuss in section 4. We will distinguish two workflows, based on whether or not there is a pre-existing lexicon that can serve as a starting point for annotation.

The first workflow (Figure 2) is intended for languages for which a digital lexical resource such as a lexical database (abbreviated in this document as 'lexicon') is already available. This may be the back-end of a public dictionary publication (whether in paper or in a digital format). If that is the case, such a lexicon can serve as the basis for corpus annotation. In other situations, the lexicon emanates from the corpus annotation process (this was the case for Corpus NGT, for instance [ ]). Such a lexicon could then in turn form the basis for a dictionary, and thus constitute a form of dissemination or public outreach. We will not be focusing on dictionary creation here, but rather focus on the lexicon as a tool for corpus annotation and as a key resource for language synthesis by avatars.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

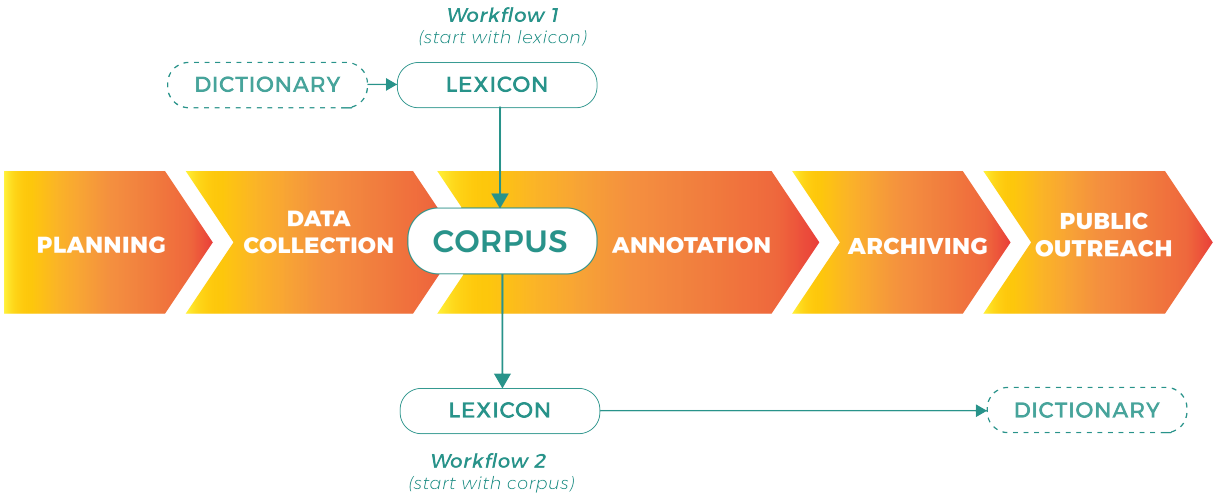

FIGURE 2: TWO TYPICAL WORKFLOWS FOR CORPUS CREATION: (1) STARTING WITH A LEXICON, (2) STARTING WITH CORPUS (SIMULTANEOUS CONSTRUCTION OF LEXICON & CORPUS)

For each stage in resource creation, we recommend identifying advisors and stakeholders within your own country that can provide input to or advice on your efforts. These can be experts

working in language institutes, but also the deaf community and the organizations around that community.

The aim for your resources should be to constitute 'FAIR' datasets (Wilkinson et al. 2016): *findable*, *accessible*, *interoperable*, and *re-usable*. Schulder & Hanke (2022) emphasize that in addition to creating data that are in compliance with those more technical demands, there are also ethical issues to be taken seriously. They discuss the 'CARE' principles (Carroll et al. 2020): there should be a *collective benefit*, the language community should have the *authority* to control data, researchers have the *responsibility* to share how data benefit the language community and their self-determination, and in general *ethical* concerns should be at the center of the whole effort. This is why we emphasize below the importance of stakeholders from not just the academic world (where the participation of deaf researchers and other deaf staff needs to be taken serious as well), but also from the language community.

We will come back to FAIR and CARE in the chapters below, in particular Chapter 6 on public outreach, where we discuss them in more detail and try to give you some handles for implementing the principles. They should both be center-stage in your project, and merit discussion and consideration during the planning phase. The fact that we discuss them later in this text rather than at the outset merely reflects practical concerns, as their meaning will be most transparent in the context of the different manuals.

1.1 STRUCTURE OF MANUALS

The five steps or phases in the process are each described in their own manual below, and each of these contain the following types of information:

- Key references and further reading
- Overall description of the phase of work
- Tools and resources
- Best practices and current standards
- Tips and considerations to keep in mind

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

2 CORPUS PLANNING MANUAL

KEY RESOURCES:

- Bowern, C. 2011. Planning a Language Documentation Project. In Peter K Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 459–481. Cambridge: Cambridge University Press.

- Chelliah, S. 2018. The design and implementation of documentation projects for spoken languages. In *The Oxford Handbook of Endangered Languages*, 147–167. Oxford University Press.

The first phase is to make a solid plan before the work begins. Much of this phase will usually be done as part of writing a proposal for funding and will involve the elements listed below. If no proposal was written beforehand or if it was not very detailed, the planning step should be given plenty of time and attention in order to have a clear path forward, prevent problems later, and make a positive impact. The plan will ideally address the following points.

- What is the **main motivation** driving the documentation? For example, is it to create a dictionary, to perform specific linguistic analyses, to understand sign language variants used in the country, to document the lives of signers (e.g. older signers), to support the learning of the sign language, etc.? Different motivations will dictate different priorities and therefore different project outcomes.
- **Timeline/Gantt chart**: a visualization of project activities plotted over time; a schematic example is shown in Figure 3.

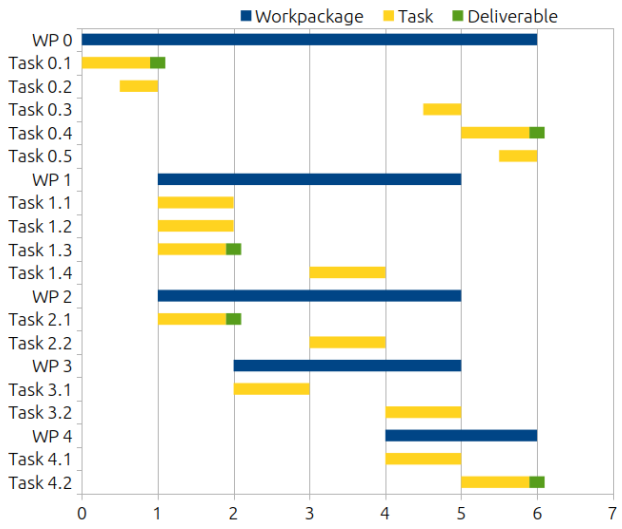

FIGURE 3: SCHEMATIC OF A GANTT CHART

(CREATED BY ZORKOW ON WIKIMEDIA. THIS FILE IS LICENSED UNDER THE CREATIVE COMMONS ATTRIBUTION-SHARE ALIKE 4.0 INTERNATIONAL LICENSE. https://commons.wikimedia.org/WIKI/FILE:WIKIMEDIA_FOUNDATION_PROJECT_GANTT_CHART.PNG)

- What are the **ultimate targets** for the corpus and lexicon to be achieved by the end of the project? Refer to the EASIER report D9.1 by Crasborn and Morgan (2022) for some qualitative and quantitative targets; e.g., number of signers per region, number of hours of recorded video, hours of tokenized video, hours of translated video, size of lexicon, etc.
- **Intermediate targets**: goals for targets that should be met by specific times (put on timeline/Gantt chart). This will show if the project is on track.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

- **Equipment, facilities, tools, software, subscriptions**, etc.: what is currently available, and what needs to be acquired? Generate a list for each phase of the project.
- Branding: create a **name and logo** for the overall project, or for distinct parts of it (e.g., corpus vs. dictionary; entire project vs. publicly-available corpus). This will help to communicate about the project to those inside and outside the project: i.e., other researchers, signing/deaf community, government agencies, public at large, etc.
- Stakeholders (1): make a **list of possible stakeholders** in each phase of the project, with suggested stakeholders at different phases shown in Figure 4. The more specific, the better; e.g., the name of the interpreters association and even the current chairperson's name rather than simply "interpreters." Project managers will want to maintain a **contact list/database** to keep information about specific stakeholders and track their interaction with the project.

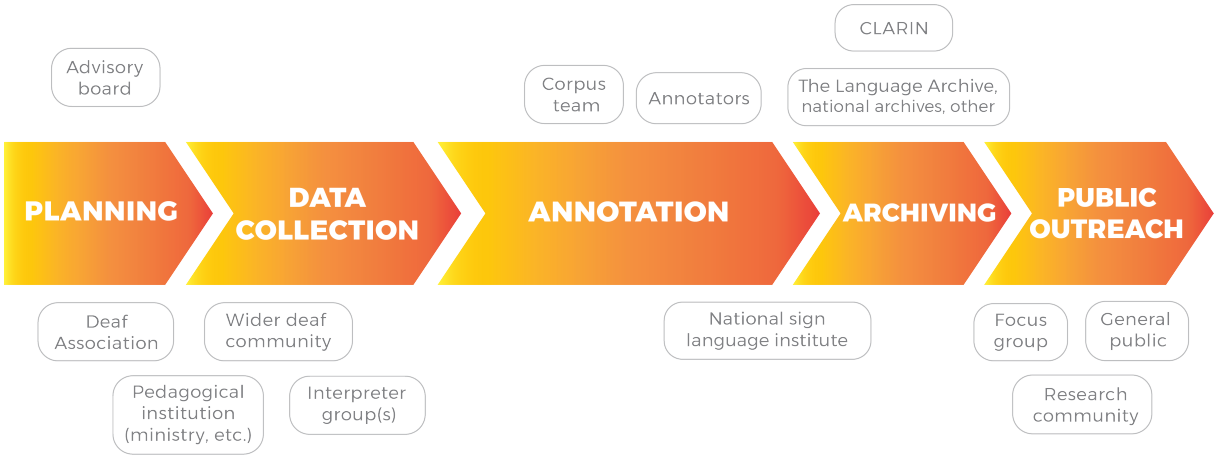

FIGURE 4: EXAMPLE OF POSSIBLE STAKEHOLDERS AT EACH PHASE

- Stakeholders (2): Entities outside of the main project team can be involved to varying degrees, from just receiving a final report, to providing occasional advice, to becoming an integral part of the project. As an example of the latter, a **language specialist advisory committee** may be formed from the deaf community to discuss lemmatization and ID-glossing choices, which would directly impact annotation. Ideally, deaf researchers would be part of the research team, of course, and involved throughout the whole lifecycle of the project.
- Stakeholders (3): **plan for involving stakeholders**; i.e., when to contact each stakeholder and sketch out (preliminary) expectations about their role/involvement, and be open to modifying the project based on stakeholder feedback.
- Ethical considerations: There are many guidelines and resources for undertaking language documentation in a responsible and ethical way. Refer to the publications posted on the website of Sign Language Linguistics Society (https://slls.eu/slls-ethics-statement/) specifically, these publications: Austin 2010; Baker 2012; Chelliah and de Reuse 2011; Crasborn 2010; Dwyer 2010; Finnish Association of the Deaf 2015; Fischer 2009; Harris et al. 2009; Johnston 2010; Morrow and Richards 1996; NIH 2009; Panda 2010; Rice 2012; Robinson 2010; Singleton et al. 2012; Singleton et al. 2015.

  Overall, these are some specific points that most projects will need to address:

  - Submit a plan to the Ethics Review Board (ERB) or Institutional Review Board (IRB) of the institution hosting the project; projects typically require **ERB/IRB approval** before they can proceed

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

- Develop a way to get **informed consent** from participants in their native language (and a way to keep it indefinitely, such as stored in metadata). Moreover, it can be important to develop an ethical way to ensure that participants are aware they cannot easily withdraw their consent, as in many cases data will be made accessible to a large audience from the initial publication.
- Develop a way to **keep sensitive data about participants secure**
- Should any data be **anonymized** in order to share it? If so, how will that be done?
- Will participants be **compensated** in any way, and if so, how?
- How will participants be informed about and included in the project?

An instrument that may be useful in some situations is a **Good Practice Agreement (GPA)**, which can be deployed to create more equal collective benefit and responsibility between the project and stakeholders (see Singleton et al. 2015) Also, take into account the *CARE Principles for Indigenous Data Governance* (https://www.gida-global.org/care), which were developed for research involving minority groups

- **Open Science**: Is the project following Open Science principles?
  - Open Access publications: build in a budget for open access and target publications that allow Open Access
  - Citizen science: are there elements that can incorporate the general public to join in part of the research?
- **Data Management Plan**: how will digital and other resources be managed? This may be dictated by funding agencies and/or universities. There are many guidelines for writing and following a data management plan. Here are some resources:
  - See The Open Handbook of Linguistic Data Management: https://doi.org/10.7551/mitpress/12200.001.0001
  - Follow FAIR principles: Findable, Accessible, Interoperable, Reusable (Wilkinson et al. 2016; see example of use in Schulder & Hanke 2022)
- **Procedures**: Working in a team and over a period of years means that important steps and information needs to be maintained with fidelity over time. Thus, as part of the regular workflow in the project, it is recommended to write procedures that can be used to train team members and to document reoccurring processes for later reference. Procedures should be dated, with an author, and storage location, and can include which tools/software were used, where files are kept, the steps to go through certain tasks etc.
- **Budget**: make a projected budget and create documents and processes to track spending throughout the project
- **Risks**: where are there likely to be problems during the project, and how will these problems be avoided, made less severe, or addressed when they arise? Write back-up plans if they are likely to be needed for some aspects of the project.
- **Outputs**: what are the planned outputs? This can include conference presentations, workshops, journal articles, website, social media, deaf community events,m and events for the general public

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

3 DATA COLLECTION MANUAL

KEY RESOURCES:

- Hanke, T. and Fenlon, J. 2002. Creating corpora: Data collection. In Hochgesang, J. & Fenlon, J. (Eds.), *Sign language corpora*. Gallaudet University Press.
- Perniss, P. 2015. Collecting and analyzing sign language data: Video requirements and use of annotation software. In Orfanidou, Eleni, Bencie Woll & Gary Morgan (eds.) *Research methods in sign language studies: A practical guide*, 53–73. John Wiley & Sons: Oxford, UK. https://doi.org/10.1002/9781118346013.ch4

Sign language corpus project almost always rely on making new recordings of language data. While creating corpora from existing sources is theoretically possible (e.g., YouTube videos, or a collection of movies in sign languages), such collections are usually not big enough and high-quality enough (but see Hou et al. 2020). Metadata about the signers are often unavailable for existing online content. Also, where data do exist, licensing restrictions can be problematic. Therefore, sign language corpus projects must rely on making new recordings of language data, usually in a controlled setting. Most large-scale projects also follow the principle of wanting to achieve a representative sample of the language, balanced for many factors related to participants and types of language use.

In general, there are five as pects of data collection to consider (summarized in Figure 5): who to include (3.1), where to film (3.2), what type of language data will be filmed (3.3), how is the filming done (3.4), and metadata about the filming (3.5).

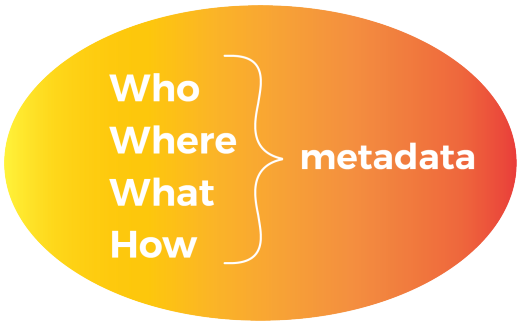

FIGURE 5: FOUR DOMAINS OF DATA COLLECTION COMING TOGETHER IN THE FIFTH, METADATA

3.1 WHO TO INCLUDE?

If the corpus is to be a proper sample of the signing community as well as be able to address research questions about language use among different groups (including controlling for such differences while focusing on other factors), then effort must be made to invite a diversity of signers by region, gender, age, ethnicity. Not only does this ensure the inclusion of all types of signers, but when social groups are carefully balanced, it makes it possible to pursue linguistic questions about language use and document the sign language in an equitable way.

For example, the corpus could be balanced on the bases of these social factors:

- Region
- Gender
- Age

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

- Socio-economic backround
- Age of sign language acquisition
- Ethnic/linguistic/religious background (depending on demographics of each country)

In trying to use project resources to capture the core of the fluent signing community for purposes of language documentation, other types of signers are naturally deprioritized. However, it is worth discussing in new documentation projects whether to include a broader swath of people, such as signers with different disabilities, hearing signers raised in a signing family with deaf members (e.g., CODAs), professional interpreters, or even L2 signers.

3.2 WHERE TO FILM?

There are two aspects here: (1) where in the country to film, and (2) in what type of physical environment to film. The first aspect will be dictated by the circumstances of each country and signing community. Are there deaf schools with different dialects, cities that are deaf hubs and attract deaf adults from the region, geological divides that correspond to different ethnic groups within the country, etc.? If so, these can be **target sites for data collection** in order to have a balanced sample.

There are other reasons to select different target sites around the country as filming locations. First, signers who are communicating with others from their region are more likely to maintain the local way of signing and not code switch into another dialect or other register. Also, being at the site greatly improve access to a range of signers, as people do not have to travel far to participate. Some corpus projects have used a **mobile studio** for this purpose (Hanke & Fenlon 2022). Yet, if the country is very small, if there are problems travelling to certain regions, or if the project involves specialized tools that cannot be transported, then it may be necessary to invite participants to travel to a single filming site.

For the second aspect, although the most natural language use happens in all types of public and private places (see section 3.3), the best quality recordings of sign language are made in **an environment that can be controlled** for light, background, and the position of the camera relative to the signer(s). See section 3.4 below for more information. Again, a mobile studio is a solution other projects have chosen to get the best possible quality while still being close to the original setting of language use.

3.3 WHAT TYPE OF LANGUAGE DATA TO FILM?

The specific language tasks or events will depend in part on the **main motivation** for the creation of the corpus (see Section 2. There are two broad types of language data: **naturalistic** and **elicited**, though these are not exclusive categories (Perniss 2015). Examples of highly naturalistic language use include spontaneous conversation between friends, interactions between a caregiver and child, work meetings, chatting over a meal, etc. Language use that is elicited is based on specific tasks, with directions given by the researcher and typically dictated by research goals. For example, if lexical variation is being studied, one method is to elicit individual signs through pictures or written text; if use of locative space is under investigation, then a participant might be asked to give directions to an interlocutor.

The most spontaneous, natural language use tends to happen at times and places where filming is difficult, whereas elicited data can be gathered in an environment with controlled lighting, cameras, backgrounds. In sign language research, clarity of movement and fine details (e.g., eyebrow furrow, finger flexion) can be crucial to capture on film (see next section

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

3.4 for more details); therefore, most corpus projects bring signers into **a controlled environment for filming**. At the same time, it is possible to increase the naturalness of the setting in various ways, such as pairing friends or acquaintances as interlocutors, using local fieldworkers to direct participants, or filming in a familiar place. For example, the BSL corpus project travelled to local deaf clubs and set up a temporary filming environment on site.

Most corpus projects plan a limited number of specific tasks to give to each participant, which vary in length and naturalness and also correspond to research goals. Table 1 shows a synopsis of the tasks collected for three different corpora (presented in Hanke & Fenlon 2022): British Sign Language (Schembri et al. 2013), Italian Sign Language (Geraci et al. 2011), and German Sign Language (described in Nishio et al. 2010).

TABLE 1 : EXAMPLE OF FILMING SESSION IN THREE CORPUS PROJECTS (PER PARTICIPANT)

Corpus:	BSL Corpus	LIS Corpus	DGS Corpus
length of filming session:	2-3 hours	3 hours	5.5 hours
tasks:	FOUR TASKS: • Signer questionnaire • brief prepared narrative • 30min free conversation • 15min narrative • lexical elicitation task	FOUR TASKS: • 40-minute free conversation • brief narrative • question elicitation session • picture naming task	TWENTY TASKS, including: • story re-telling from a prompt • discussion on given topic with mediator • free conversation • etc.

Below are examples of **language tasks** that have been used in other corpus projects. A more comprehensive list of tasks used in corpus projects can be found in the online Dataset Compendium (Kopf et al. 2022b) and in a previous EASIER report D9.1 (Kopf et al. 2022a).

- participant explains the origin of their sign name
- participant describes their educational background
- participant is asked their opinion about a topic currently being debated
- participant is asked to tell a deaf culture joke
- participant watches a short film or animation, then explains what happened (films used: *Canary Row*, *Frog Where are You?*, *The Snowman*, *Pear Film*)
- participant reads a short vignette or story and explains what happened; e.g., the book *Frog Where are You?* has been used in this task
- participant and partner both have a similar picture; one person explains to the other person what is on their picture, or they have to discover the differences between pictures without looking at the one provided to the other person
- two (or more) participants talk freely in a group for about 30 minutes while the fieldworker waits outside
- two participants are given calendars with full schedules and asked to negotiate a time and date to meet
- participant is asked their sign for a concept presented in a picture; may also be asked to use the sign in a sentence

3.4 HOW TO FILM?

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

Another ingredient in data collection is how the filming is actually done. This involves so many factors that is recommended to run a **pilot of the entire process** in advance, from inviting a participant to be filmed, getting permission, collecting demographic data ('metadata', see 3.5), setting up all the equipment, doing the filming with all tasks, and managing the video output and participant permissions with filenames and storage. This will expose any existing weaknesses in the process, reveal the administrative work required, and give the project managers feedback to make projections for data collection targets in the project timeline (see Gantt chart above). In addition, a pilot will help with writing procedures (or filming them in sign) for future use by team members (e.g., in training new fieldworkers or coders).

Here is a list of things to consider:

- Cameras
  - See details in Perniss (2015: 60–61) and Hanke & Fenlon (2022: 35–38)
  - How many? What position?
    - Assuming a set-up for pairs of signers, it depends somewhat on the budget, practical considerations, and research questions
  - Different options are possible; here are two possible scenarios:
    - **Scenario with 3 cameras**: two for a frontal view for each signer in a pair, one wide-angle lens camera with a top view (at an angle; not necessarily fully overhead) to capture all participants in a single frame
    - **Scenario with 4-5 cameras**: two for a frontal view for each signer in a pair, one to capture all participants in a single frame from the side, and 1-2 fully overhead (pointing downward) to capture movement around the signers.
  - Type of camera(s)
    - "meet the expectations of future researchers" (H&F 2002: 35) by using best quality available within the project budget
    - As of 2022, avoid anything below 2K resolution (i.e., "full HD"); a target of 4K resolution can allow later digital zoom into details
    - Look for high frame rates – ideally 100/120 frames per second (fps) because 50/60 fps can cause motion blur; standard framereates of 25/30 fps are not good enough for capturing signing movements
    - Tight budget? Minimum consumer level product in 2022: full-HD camera with 50/60 fps (H&F 2022) and use of good lighting
- Lighting
  - Getting a good lighting kit is worth the investment because it allows a camera to capture more fine-grained detail, which can offset lower frame rates
  - Lighting should provide full but not blinding light (may need to use a diffuser)
  - Position to avoid deep shadows but also allow some contrast

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

- Some lighting rigs can create heat, so consider filming conditions in the summer or in a hot climate

- Background
  - Plain and even in appearance; not a patterned background
  - Options (1): hanging cloth, screen, painted room
  - Options (2): blue or green screen for later video processing (though this is not required for the tools being developed in EASIER)
- Signers
  - Wear plain clothes that contrast with skin
- Furniture / objects in the room
  - Chairs that don't swivel (to preserve camera angle)
  - Chairs that don't have armrests (to allow freedom of movement in signing)
  - Elicitation materials out of the way (not on a table); may be presented on monitors on the floor or a monitor/screen/whiteboard/paper off-camera
  - Low tables to the side of participants or fieldworker; only for necessary objects (keyboard, mouse, drinks)
- Cue to signal both beginning of filming and to synchronize multiple cameras; e.g., lights blink, clapperboard, hands clapping
- Metadata
  - May want to use a form, checklist, and/or questionnaire to make sure all metadata captured during data collection
  - Clapperboard with blackboard/whiteboard at the beginning/end of filming can be used for some metadata (date/time, location, participant names or codes
- Other
  - Keep track of time to allow participants to take breaks
  - Plan how to stop and start cameras: remote controls? Order of cameras?
  - Try to run all cameras from electricity, but have multiple charged back-up batteries in case electricity is a problem
  - Mobile studio? Filming for many hours at host site will use a lot of electricity; consider compensating the host for the use of their facilities

3.5 METADATA

In the Archiving Manual (Section 5), we will come back to the need for standardised metadata. **Metadata** refers to all the information about episodes of signing captured on video: who, what, where, when, and how. This is covered in more depth in Section 5 below.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

To create these metadata, it is vital that information is collected about the sessions being recorded and about the signers at the time of filming.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

4 ANNOTATION MANUAL

RELEVANT DOCUMENTS FROM THE EASIER PROJECT:

- Kopf, M., Schulder, M., Hanke, T., and Bigeard, S. 2022. *Specification for the Harmonization of Sign Language Annotations* (EASIER deliverable D6.2). https://doi.org/10.25592/UHHFDM.9841

- Crasborn, O., and Morgan, H. 2022. *Definition of minimal contents of datasets for participation* (EASIER deliverable D9.1). Radboud University.

OTHER KEY RESOURCES:

- Hodge, G., and Crasborn, O. 2022. Good practices in annotation. In J. Hochgesang & J. Fenlon (Eds.), Sign language corpora. Gallaudet University Press.

- Johnston, T. 2019. Auslan Corpus Annotation Guidelines (p. 105). https://auslan.org.au/about/annotations/

- Published annotation guidelines for other sign languages;
    - Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. 2022. Public DGS Corpus: Annotation Conventions. Version 4.1. Hamburg University. https://doi.org/10.25592/uhhfdm.822
    - Crasborn, O., Zwitserlood, I., Kooij, E. van der, Bank, R., and Ormel, E. 2020. Annotation conventions for the Corpus NGT. Version 4. https://www.ru.nl/publish/pages/1013556/corpusngt_annotationconventions_v4_1.pdf

- Schembri, A., and Crasborn, O. 2010. Issues in creating annotation standards for sign language description. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. M. Ruiz, & A. Schembri (Eds.), Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (pp. 212–216). ELRA. https://www.sign-lang.uni-hamburg.de/lrec/pub/10050.html

- Cormier, K., Crasborn, O., and Bank, R. 2016. *Digging into Signs: Emerging Annotation Standards for Sign Language Corpora* (E. Efthimiou et al, Ed.; pp. 35–40). ELRA. https://www.sign-lang.uni-hamburg.de/lrec/pub/16015.html

In order to be accessible to computers – that is, *machine-readable* – video data of free conversation, narratives, interaction, and other tasks must be **annotated**. This is what allows computers to search for signs or other units. Eventually, language technologies as developed in EASIER will facilitate automatic or machine-supported annotation, but currently the initial annotations will come from human work. This is a very slow process: count on one hour of recording requiring at least 250 hours of work for providing sentence-level translations and glossing of manual signs. Annotations on these two levels are essential for making your sign language corpus accessible to sign language technologies, along with a lexical database that contains a video-recorded citation form of signs, possible meanings (translation equivalents), and a transcription of the sign at the level of phonetics or phonology.

The present section refrains from presenting detailed manual for annotation, as these have been published for various corpora already. All examples listed above would form a good

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

starting point. In this section, we make some general comments and recommendations that can help you create high-quality annotated data.

The graphic rapresentation of the stages in corpus development in the figures above suggests that each step is a finite process, after which one transitions to the next stage. This is especially deceptive for annotation, for two reasons. First, as annotation is so time-consuming, it is not uncommon for annotation to last long after the initial data were deposited in an archive. Newer annotations in turn will also have to be archived, so that also archiving can be a repetitive process. In the Corpus NGT for instance, four subsequent releases of annotations have been created, three of which have been incorporated in The Language Archive so far. Second, annotation is typically also a cyclic process in that old annotations will be revised following new insights and changes in annotation conventions over time. This has been the case for the four Corpus NGT annotations for instance, and each release is thus accompanied by an update to the annotation guidelines and an overview of what changed since the last release.

Although a rich transcription of signs would include not just manual but also non-manual annotations, because nearly all of the lexicon is produced by the hands with a limited role for non-manuals, transcribing or glossing lexical signs is the biggest contribution one can make to making a corpus machine-readable: what the hands do is most variable and least predictable.

The form of signs (phonetic description) can be transcribed using a system like HamNoSys (Hanke 2004) or other types of coding. Whatever transcription system is used, this is done once per sign rather than transcribing the parts of a sign each time it appear: at the level of the type rather than at the level of the token. Thus, glosses need to be '**ID-glosses**' (Johnston 2010), that is, not so much translations of the sign in context, but unique identifiers that link a corpus to a lexical database, where the phonetic description is stored (as well as other lexical information). Although most glosses will be composed of the generic meaning of the sign and possibly an index value to tell apart different signs that share the same meaning (e.g., 'HOME-01', HOME-02'), some codes will also be necessary to classify the content of signing that is not found in the lexical database in a standard form. This includes classifiers, fingerspelling strings, and other partly-lexical or non-lexical manual actions (Johnston & Schembri 1999). We recommend studying some of the larger existing annotation guidelines, like those for the Auslan corpus (Johnston 2019) or Corpus NGT (Crasborn et al. 2015), and to follow the recommendations for standardisation of glosses produced by the EASIER project.

Aside from linguistic considerations, a key requirement for glosses is that they are easily parsable for the computer ('**machine-readable**'). Make sure not to re-use the same delimiter characters (like hyphens and colons) for different purposes. For example, if you use hyphens as spacing in multi-word expressions like AROUND-THE-CLOCK, use a different symbol to separate this expression from other parts of the gloss (e.g. $NON-MANUAL:TO-KISS instead of $NON-MANUAL-TO-KISS). The resulting string should be parsable by regular expressions – something that computer scientists will be able to help you achieve. Following existing conventions for the larger corpora will help you attain this as well, and make sure to check and follow Deliverable 6.2 of EASIER (Kopf et al. 2022a).

Glosses are typically time-aligned one by one in software like ELAN or iLex, although it is conceivable that sequences of glosses are produced in larger sentence-level annotations. The latter would speed up the annotation process, at the expense of locating specifically where each manual action starts and ends within a sentence. Current technologies may soon be able to help to automatically align glosses based on such annotations. We recommend finding out what possibilities are available at the moment that you start the annotation process. This may involve liaising with language technology institutes in your country, as well as setting up support from a software developer or language technologist at your own institute.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

Finally, a general recommendation is to document well what you have annotated. This means releasing annotation conventions that accompany the corpus. When the corpus expands, this also means releasing a new version of the corpus with accompanying documentation and updated annotation conventions. It is a good idea to keep older versions accessible. A single web page with an overview of all your releases can be helpful (or a PDF added to the archive where you deposit your annotations along with the videos).[1]

[1] For example, see the releases of the the DGS corpus listed here: https://www.sign-lang.uni-hamburg.de/meinedgs/landing/corpus_en.html.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

5 ARCHIVING MANUAL

KEY RESOURCES:

- Berez-Kroeker, A. L., McDonnell, B., Koller, E., and Collister, L. 2022. *Open Handbook of Linguistic Data Management.* MIT Press Open; it contains the following chapters on sign language data, but other chapters will also be of interest:
    - Crasborn, O. 2022. Managing Data in Sign Language Corpora. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister (Eds.), The Open Handbook of Linguistic Data Management (pp. 463–470). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0044
    - Palfreyman, N. 2022. Managing Sign Language Data from Fieldwork. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), The Open Handbook of Linguistic Data Management (p. 11). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0026
    - Hochgesang, J. A. 2022. Managing Sign Language Acquisition Video Data: A Personal Journey in the Organization and Representation of Signed Data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), The Open Handbook of Linguistic Data Management. The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0035
    - Hou, L., Lepic, R., and Wilkinson, E. 2022. Managing Sign Language Video Data Collected from the Internet. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), The Open Handbook of Linguistic Data Management (pp. 471–480). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0045
- Smith, R. T., Willoughby, L., and Johnston, T. 2022. Integrating Auslan Resources into the Language Data Commons of Australia. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources (pp. 181-186). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22017.html
- Kuder, A., Wójcicka, J., Mostowski, P., and Rutkowski, P. 2022. Open Repository of the Polish Sign Language Corpus: Publication Project of the Polish Sign Language Corpus. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources (pp. 118–123). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22010.html

Data can be archived in many ways, making them 'findable' and 'accessible' (see the FAIR principles in Section 6). Although any server might host the data, it is advisable to also consider depositing a copy of the data in a national or international archive. There are various repositories that are dedicated to linguistic data, such as the Linguistic Data Consortium (LDC) in the USA, the European Language Resources Association (ELRA) in Paris, or the The Language Archive (TLA) of the Max Planck Institute for Psycholinguistics in Nijmegen. In the context of European projects on language resources such as CLARIN, smaller centers have also been identified and/or strengthened that cover national data or also services to data depositors from other countries. For some types of resources, the Endangered Languages Archive (ELAR) may also be a suitable location. Finally, Zenodo (https://zenodo.org) is an open

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

data repository for all types of scientific datasets; it is open access, free of charge, has long-term support, and can issue a stable DOI (digital object identifier) to uniquely identify your dataset(s).

In all cases, metadata are vital: these are the cataloguing information by which datasets can be found and sessions within a dataset can be identified. Metadata are thus 'data about the data' and falls into one of two categories:

(1) **participant metadata**: information about each signer that is gathered, often in a questionnaire process at some point, and can be in a written form or on camera. It will include basic information like name, age, educational background, etc., or be a more detailed profile depending on the research focus of the project.

(2) **session metadata**: information about the filming event itself: when, where, who (participants, moderator, etc.), how (camera specs and setup), etc.

Metadata serve several purposes. First, they must be connected with the annotated data to enable research, such as investigating differences in language use by region, age, etc. Second, metadata can be used during the project as a means of quickly locating certain videos for any purpose: e.g., looking up all signers from a particular region, or all videos of overhead cameras used in a particular month, or all sessions with a specific person as the mediator, etc. Third, it is a key part of archiving, which allows future users to later know details about each video: who is in the video, when was it filmed, what happens in it, how it was filmed, etc.

When archiving data, there are two steps to providing metadata: The first step is providing the metadata that the archive itself requests for each submission, which is mostly information common to all kinds of datasets like its name and authors (although some archives might also request information specific to language data). It is usually provided by filling out a web form while uploading the data. The second step is to include metadata files in your dataset that cover relevant information which the archive did not ask about. It is best if these files are in a standardized, machine-readable format such as CMDI, which is designed specificially for language data. For instance, The Language Archive has specified a specific CMDI metadata profile for sign language data ('lat-SL-session') incorporating some of the categories above.[2] Corpora that make use of CMDI metadata are for example the DGS corpus (see discussion in Hanke, 2021) and the Corpus NGT (available from The Language Archive, https://hdl.handle.net/1839/b0c69aeb-222a-41df-b7c5-979001b635b3).

When making metadata public, you also need to keep in mind how this affects the privacy of the participants. The metadata that archives request are always public, even when access to the data is restricted, to make the data 'findable' (the F in FAIR) by both humans and web crawlers. The metadata you provide as files may be public or restricted, depending on what you chose. Thus, privacy-sensitive data that are important for use of the corpus by researchers should only ever be included in the archive as restricted-access documents, while the metadata that are public can only contain information that signers have given consent for.

[2] See the CLARIN Component Registry, https://catalog.clarin.eu/ds/ComponentRegistry, or go directly to the profile: http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1417617523856. This profile includes some of the values emerging from European workshops of sign linguists, see e.g. Crasborn et al. (2007) and Crasborn & Hanke (2003), which include such data categories as 'sign language experience: exposure age'; 'sign language experience: acquisition location'; 'education: education model'; 'family: mother: primary communication form'; 'interpreting: source'; 'interpreting: target'; and others.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

This consent is often obtained at the same time that participants fill out an information form to obtain the participant metadata mentioned above. The details of the **consent form** determine the access configuration for the various media and annotation files, but are typically not stored in the metadata fields themselves. An argument in favor of doing so is that informed consent is often complex: signers may consent to use for research and sharing through an archive, but perhaps not for use in teaching or in printed publications (see e.g. Hanke et al. 2010).

Finally, one should consider the conditions under which others can use the data. The user license for the data should match the consent obtained from the signers. Public licenses like those of Creative Commons (https://creativecommons.org) offer a variety of options that may suffice, depending on the content of your recordings. Such public use licenses may not be commensurate with the CARE principles discussed above, though. Possibly, the archive where you deposit the data also offers specific end user agreements containing more specific licensing conditions. It can also be a good idea to consult with data at your own institute or in your country.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

6 PUBLIC OUTREACH MANUAL

KEY RESOURCES:

- Adam, R. 2015. Dissemination and transfer of knowledge to the Deaf community. In Orfanidou, Eleni, Bencie Woll & Gary Morgan (eds.) *Research methods in sign language studies: A practical guid*e, 41–52. John Wiley & Sons: Oxford, UK. https://doi.org/10.1002/9781118346013.ch3

- Schulder, M. and Hanke, T. 2022. How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages. In *Proceedings of the 13th Language Resources and Evaluation Conference*. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.18

- Papers published at sign-lang@LREC: Workshop Series on the Representation and Processing of Sign Languages at LREC (International Conference on Language Resources and Evaluation). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec

While this manual appears last, it is not because the topics here can wait until the end. It is placed here because language data is in the best condition to share with others toward the end of the project. However, *interaction between the corpus builders and various stakeholders should be continual throughout the project*. Recall that the manual for corpus planning (Section 2) lists the identification of **stakeholders** as part of the process, from beginning to end. Stakeholders include any groups, organizations, or individuals who will benefit or be impacted by this research. In this manual, a brief synopsis of the principles regarding community interaction is given, followed by an overview of reaching various audiences.

6.1 PRINCIPLES REGARDING OUTREACH TO THE PUBLIC

In the past few decades, strands of thought from academia, funding agencies, nonprofit and aid organisations, and governments have come together to create new expectations about the relationships between various actors (researchers, volunteers, etc.), the communities they study or work with, and the general public. These are some of the main ideas relevant to sign language documentation:

- Researchers need more access to each other's work in order to avoid wasting resources and to make scientific advancements more effectively and efficiently
- The general public needs to have a better understanding about what scientists are doing in their research
- Research involving indigenous communities should be undertaken in consultation with those communities
- Deaf people or SLPs (Sign Language Peoples) "share significant life experiences, including patterns of oppression" with indigenous peoples (Batterbury et al 2007; see also Bone et al. 2021)

Guiding principles have been drawn up to address these ideas. **FAIR** (Wilkinson et al. 2016) addresses the quality and accessibility of data, while **CARE** (Carroll et al. 2020; Research Data Alliance International Indigenous Data Sovereignty Interest Group 2019) addresses how principles should be enacted when the data involves a minority group. These are defined

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

further in Table 2. For a specific example of how these principles were enacted in a corpus project, see Schulder & Hanke (2022).

TABLE 2 : FAIR & CARE PRINCIPLES FOR COLLECTION & MANAGEMENT OF DATA

FAIR Principles	CARE Principles
Findable: Data should be easy to find for both humans and machines. This requires globally unique and persistent identifers which are indexed in searchable resources and associated with rich metadata.	**Collective Benefit**: Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.
Accessible: Users need to know how to access (meta)data, possibly including steps for authentication and authorization. Access should be defined by metadata and use free and open protocols. Even when data is no longer available, its metadata should be.	**Authority to Control**: Indigenous Peoples' rights and interests in Indigenous data must be recognized and their authority to control such data be empowered.
Interoperable: Data usually needs to be integrated with other data and interoperate with applications for analysis, storage and processing. (Meta)data should use well-defined knowledge representation formalisms, open controlled vocabularies and include qualified references to other (meta)data.	**Responsibility**: Those working with Indigenous data have a responsibility to share how this data is used to support Indigenous Peoples' self determination and collective benefit.
Reusable: Data and metadata should be well-described so they can be re-used in different settings. They should have a clear license, detailed provenance information and meet domain-relevant community standards.	**Ethics**: Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

These principles are also a component of the broader movement for **Open Science**, which encompasses a variety of efforts to make scientific research and its dissemination accessible to all levels of society, amateur or professional. For more information, see the Open Science Training Handbook (Bezjak et al. 2018, or its more readable version here: https://book.fosteropenscience.eu/). This dovetails with emerging focus on **science communication** as a set of skills and practices that can bridge the gap between researchers within their own fields and everyone else, from other scientists to the general public (Fischhoff & Scheufele 2013).

Note that following the guidelines in Section 5 to **archive** the project data ensures a level of accessibility, and we recommend that dataset be made available to the general public, as much as possible. One way to do this is by creating a dedicated web platform that allows users to search in translations and other annotations, as has been done for the LSFB corpus (Sinte et al. 2015) and the DGS corpus (Jahn et al. 2018; Isard and Reiner 2022), for instance.

6.2 TYPES OF OUTREACH AND DISSEMINATION

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

Who should you reach out to? As explained in Section 2, the target audiences for dissemination should be the same as the list of **stakeholders** created during the project planning phase, and including new contacts gathered during the project; e.g., various social groups, organizations, and even specific individuals. Also, of course, the scientific community will be part of any outreach and dissemination efforts.

Depending on the material to share and the target audience, dissemination happens in different venues and is either more passive (e.g., a website) or more active (e.g., giving a presentation at a deaf school). Here are the overall types of dissemination:

- **Broadcast dissemination**: passive posting of information from main research center/lab/office for others to find
  - Websites:
    - Website describing the project: what, when, funding, news, contact info, etc.
    - Online public corpus for viewing and searching the data
    - Online private access corpus (e.g., for researchers via registration)
    - Annotation conventions and/or coding manual, specific to dataset release/version
    - Archives (see Section 5)
      - Full corpus archives (with video), one for each release/version
      - Corpus *transcript* archiving, specific to dataset release/version
      - Metadata
    - Research materials posted online: publications, presentations, posters, videos
    - Social media
- **Embedded dissemination**: participation in pre-planned events with a wider audience than local research group (lectures, activities, panel discussions, citizen science events, etc.)
  - Academic events: conferences, workshops, lecture series, invited talks, etc.
  - Science fairs
  - Community festivals
- **Targeted dissemination**: scheduled outreach to specific audiences, usually at their sites
  - Deaf schools: presentation, activities
  - Classroom presentations in other schools, for all ages
  - Presentation to an organization: e.g., Deaf club, CODA group, interpreter association, deaf educators, ministry of education, etc.

6.3 CASE STUDIES: OUTREACH AND EXTENSION OF CORPUS RESOURCES

There are relatively few accounts about efforts in public outreach and managing existing corpus resources compared with doing new documentation, annotation methods, or even linguistic discoveries based on a corpus. Therefore, project managers are encouraged to reach

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

out to other sign language corpus managers to find out what they have done or are currently doing at later stages in the life of a corpus. They may have helpful feedback and novel ideas for engaging the public, as well as cautionary advice.

Corpus managers are also encouraged to attend the **Workshop Series on the Representation and Processing of Sign Languages** at LREC (International Conference on Language Resources and Evaluation), known also by the tag **sign-lang@LREC**, which is held every two years. This is an important venue for meeting other people and teams working on sign language documentation.

Here are three recent case studies, presented at the most recent LREC, which highlight different aspects of how corpus data may need to be managed even into the more mature stages, with respect to public interaction and long-term access.

- First, for the DGS-Korpus project, Jahn et al. (2022) review where their project was more and less successful in achieving an interactive exchange between corpus managers and the language community. Among other discoveries, they highlight that **focus groups** of community experts became crucial to effective communication about the project. They recommend this as a standard best practice for other corpus projects.
- Second, in the corpus of PJM (Polish Sign Language), Kuder et al. (2022) describe how their project team worked to satisfy the highest level of access possible for both the general public and for researchers by offering **two separate repositories**. They explain the challenges of having to manage sensitive material on video, as well as the best structure and visual environment for the website, and the structure of data in the researcher repository in order for it to be usable by linguists.
- Third, Smith et al. (2022) describe the integration of the Auslan corpus into a larger network within Australia called the Language Data Commons of Australia (LdaCA; https://www.ldaca.edu.au). The Auslan corpus is the oldest digital sign language corpus in the world, and this paper illustrates two principles: (1) a corpus can continue to undergo updates and improvements indefinitely as long as there are linguists and funding to do so, and (2) at some point, it may be prudent to think about **very long-term archiving** depending on the current location and status of a corpus. This example shows that one way this can be done is through a national framework to manage language resources.

These case studies reveal that the work of corpus management does not end when the initial annotation targets have been reached, and that integration “downward” into the community and “upward” into a super-structure of other language resources can result in a truly lasting contribution with broad societal impact. See also the comments on annotation above: annotation is typically both a long-lasting and a cyclic process, that requires long-term attention.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

7 CONCLUSION

This report has aimed to provide an overview of a sign language documentation project from beginning to end, in five overall steps or phases, with connections to key resources and recommendations at each step. Language documentation — resulting in a corpus, lexical index/database, and potentially a dictionary — takes time, funding, institutional support, the input of various experts, and interaction with many different stakeholders.

Yet, each project is also unique; the main goals of each project will vary, as will the specific groups involved and their expectations, and the resources available. We hope that this guide will lay out a path to achieving a high quality sign language corpus that suits the needs of the community as well as researchers, and prepares the way for under-resourced languages to be able to benefit from functional workflows of language technologies in the future, such as those being developed in the EASIER project.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

REFERENCES

[1] Adam, R. 2015. Dissemination and transfer of knowledge to the Deaf community. In Orfanidou, Eleni, Bencie Woll and Gary Morgan (eds.) *Research methods in sign language studies: A practical guide*, 41–52. John Wiley & Sons: Oxford, UK. https://doi.org/10.1002/9781118346013.ch3

[2] Austin, P. K. 2010. Communities, ethics and rights in language documentation. In P. K. Austin (ed.), *Language documentation and description* (Vol. 7), 34-54. London: SOAS University of London.

[3] Baker, A. 2012. Ethics issues in sign language acquisition data archiving. Ms., University of Amsterdam. IPROSLA: Integrating and Publishing Resources of Sign Language Acquisition.

[4] Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P.L., Görögh, E., Helbig, K., Kramer, B., Labastida, I., Niemeyer, K., Psomopoulos, F. and Ross-Hellauer, T., 2018. Open Science Training Handbook (1.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.1212496 and https://book.fosteropenscience.eu/

[5] Bowern, C. 2011. Planning a Language Documentation Project. In Peter K Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 459–481. Cambridge: Cambridge University Press.

[6] Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez- Lonebear, D., Rowe, R., et al. 2020. The CARE principles for Indigenous data governance. *Data Science Journal*, 19(43):1–12. DOI: http://doi.org/10.5334/dsj-2020-043.

[7] Chelliah, S. 2018. The design and implementation of documentation projects for spoken languages. In *The Oxford Handbook of Endangered Languages*, 147–167. Oxford University Press.

[8] Chelliah, S. L. and de Reuse, W.J. 2011. Handbook of Descriptive Linguistic Fieldwork. Dordrecht: Springer. [Chapter 6: Fieldwork Ethics: The Rights and Responsibilities of the Fieldworker]

[9] Cormier, K., Crasborn, O., and Bank, R. 2016. *Digging into Signs: Emerging Annotation Standards for Sign Language Corpora* (E. Efthimiou et al, Ed.; pp. 35–40). ELRA.

[10] Crasborn, O A., Mesch, J., Waters, D., Nonhebel, A., Van der Kooij, E., Woll, B., and Bergman, B. 2007. Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics 12* (4): 535-562.

[11] Crasborn, O., Zwitserlood, I., Kooij, E.van der., Bank, R., and Ormel, E. 2020. Annotation conventions for the Corpus NGT. Version 4. https://www.ru.nl/publish/pages/1013556/corpusngt_annotationconventions_v4_1.pdf

[12] Crasborn, O. 2010. What does 'informed consent' mean in the Internet age? Publishing sign language corpora as open content. Sign Language Studies 10(1): 276-290.

[13] Crasborn, O. 2022. 39: Managing Data in Sign Language Corpora. *The Open Handbook of Linguistic Data Management*: 463–470. https://doi.org/10.7551/mitpress/12200.003.0044

[14] Crasborn, O., and Morgan, H. 2022. *Definition of minimal contents of datasets for participation* (EASIER deliverable D9.1). Radboud University. https://www.project-easier.eu/wp-content/uploads/sites/67/2022/08/EASIER_D9.1_Definition-of-minimal-contents-of-data-set-for-participation_v1.0.pdf

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

[15] Dwyer, A. M. 2010. Models of successful collaboration. In L. A. Grenoble & L. N. Furbee (eds.), Language documentation. Practice and values, 193-212. Amsterdam Philadelphia: John Benjamins Publishing Company.

[16] Finnish Association of the Deaf. 2015. Chapter 6: Best Practices and Challenges in Sign Language Work – in Working Together: Manual for Sign Language Work within Development Cooperation.

[17] ßConcerns. In Taiwan Sign Language and Beyond. 2009, James H-Y. Tai and Jane Tsay, eds., 1-19. Chia-Vi, Taiwan: The Taiwan Institute for the Humanities, National Chung Cheng University.

[18] Hanke, T. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In Streiter, Oliver, Vettori, Chiara (eds): *LREC 2004, Workshop proceedings: Representation and processing of sign languages*. Paris : ELRA, 2004, 1–6

[19] Hanke, Thomas. 2021. *Persistent Identifiers and Metadata for the Public DGS Corpus*. Project Note AP06-2021-01, Hamburg University. https://doi.org/10.25592/UHHFDM.10219.

[20] Hanke, T. and Fenlon, J. 2022. Creating corpora: Data collection. In Hochgesang, J. & Fenlon, J. (Eds.), *Sign language corpora*. Gallaudet University Press.

[21] Hanke, Thomas, Hong, Sung-Eun, König, Susanne, Langer, Gabriele, Nishio, Rie, & Rathmann, Christian. 2010. *Towards Fair Licences for Data from the DGS Corpus Project*. Poster presentation, 4th Sign Language Corpora Network Workshop: Exploitation (SLCN4), Berlin, Germany, 3-4 Dec 2010, Berlin. https://doi.org/10.25592/UHHFDM.1885

[22] Harris, R., Holmes, H.M., and Mertens, D.M. 2009. Research ethics in sign language communities. *Sign Language Studies 9*(2): 104–131. https://doi.org/10.1353/sls.0.0011

[23] Hodge, G., and Crasborn, O. 2022. Good practices in annotation. In J. Hochgesang and J. Fenlon (Eds.), *Sign language corpora*, 46–89. Gallaudet University Press.

[24] Hou, L., Lepic, R., & Wilkinson, E. 2020. Working with ASL Internet Data. *Sign Language Studies*, *21*(1), 32–67. https://doi.org/10/ghmbkz

[25] Isard, A. and Konrad, R., MY DGS–ANNIS: ANNIS and the Public DGS Corpus. In *Proceedings of the 13th Language Resources and Evaluation Conference*. https://www.sign-lang.uni-hamburg.de/lrec/pub/22034.html

[26] Jahn, E., Konrad, R., Langer, G., Wagner, S. and Hanke, T. 2018. Publishing DGS corpus data: Different formats for different needs. *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC.* https://www.sign-lang.uni-hamburg.de/lrec/pub/18018.html

[27] Jahn, E., Khan, C. & Herrmann, A. 2022. Outreach and Science Communication in the DGS-Korpus Project: Accessibility of Data and the Benefit of Interactive Exchange between Communities. https://www.sign-lang.uni-hamburg.de/lrec/pub/22035.html

[28] Johnston, T. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics, 15(1), 104–129. https://doi.org/10.1075/ijcl.15.1.05joh

[29] Johnston, T. 2019. Auslan Corpus Annotation Guidelines. https://auslan.org.au/about/annotations/

[30] Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. 2019.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

Public DGS Corpus: Annotation Conventions (p. 27). Hamburg University. https://doi.org/10.25592/uhhfdm.822

[31] Kopf, M., Schulder, M., and Hanke, T. 2021. *Overview of Datasets for the Sign Languages of Europe* (EASIER deliverable D6.1). https://doi.org/10.25592/uhhfdm.9561

[32] Kopf, M., Schulder, M., Hanke, T., and Bigeard, S. 2022a. *Specification for the Harmonization of Sign Language Annotations* (EASIER deliverable D6.2). https://doi.org/10.25592/UHHFDM.9841

[33] Kopf, M., Schulder, M., & Hanke, T. 2022b. The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources. *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, 102–109.

[34] Kuder, A., Wójcicka, J., Mostowski, P., and Rutkowski, P. 2022. Open Repository of the Polish Sign Language Corpus: Publication Project of the Polish Sign Language Corpus. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources (pp. 118–123). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22010.html

[35] LSA Ethics Statement. http://lsaethics.wordpress.com/2008/07/

[36] Morgan, H. and Crasborn, O. 2022. *An Overview of Resources in the Making* (EASIER deliverable D9.2). Radboud University. https://www.project-easier.eu/wp-content/uploads/sites/67/2022/08/EASIER_D9.2_An-overview-of-resources-in-the-making_v1.0.pdf

[37] Morrow, V. and Richards, M. 1996. The ethics of social research with children: An overview 1. *Children & society 10*(2): 90–105. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1099-0860.1996.tb00461.x

[38] NIH. 2009. Research Involving Individuals with Questionable Capacity to Consent: Points to Consider. http://grants.nih.gov/grants/policy/questionablecapacity.htm

[39] Palfreyman, N. 2022. Managing Sign Language Data from Fieldwork. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), The Open Handbook of Linguistic Data Management (p. 11). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0026

[40] Panda, Sibaji (ed.). 2010. Technology and Ethics in Sign Language Research. Nijmegen: Ishara Press. (DVD)

[41] Perniss, P. 2015. Collecting and analyzing sign language data: Video requirements and use of annotation software. In Orfanidou, E., Woll, B. & Morgan, G. (eds.) *Research methods in sign language studies: A practical guide*, 53–73. John Wiley & Sons: Oxford, UK. https://doi.org/10.1002/9781118346013.ch4

[42] Research Data Alliance International Indigenous Data Sovereignty Interest Group. (2019). CARE principles for indigenous data governance. *https://www.gida-global.org/care*

[43] Rice, Keren. 2012. Ethical Issues in Linguistic Fieldwork. In N. Thieberger, Nicholas, (ed.), The Oxford Handbook of Linguistic Fieldwork, 407-429. Oxford/New York: Oxford University Press.

[44] Robinson, L. C. 2010. Informed consent among analog people in a digital world. Language & Communication 30(3): 186-191.

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union

https://doi.org/10.1016/j.langcom.2009.11.002

[45] Schembri, A., and Crasborn, O. 2010. Issues in creating annotation standards for sign language description. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. M. Ruiz, & A. Schembri (Eds.), 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (pp. 212–216). ELRA. https://www.sign-lang.uni-hamburg.de/lrec/pub/10050.html

[46] Schulder, M. and Hanke, T. 2022. How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages. In *Proceedings of the 13th Language Resources and Evaluation Conference*. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.18

[47] Singleton, J.L., Jones, G. and Hanumantha, S. 2012. Deaf Friendly Research? Toward Ethical Practice in Research Involving Deaf Participants. Deaf Studies Digital Journal 3. https://doi.org/10.1177/1556264614540589

[48] Singleton, J. L., Martin, A. J. and Morgan, G. 2015. Ethics, Deaf-Friendly Research, and Good Practice When Studying Sign Languages. In E. Orfanidou, B. Woll and G. Morgan (eds.), Research Methods in Sign Language Studies: A Practical, John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781118346013.ch1

[49] Sinte, A., De Clerck, C., Fonzé, S., Sanchez, S., Raes, G., Meurant, L. 2015. Corpus LSFB (French Belgian Sign Language): Current annotation conventions compared to the "Digging into signs" suggestions. Poster presented at *Digging into Signs Workshop*. https://bslcorpusproject.org/wp-content/uploads/LSFB_Sinte-et-al._Poster.pdf

[50] Smith, R. T., Willoughby, L., and Johnston, T. 2022. Integrating Auslan Resources into the Language Data Commons of Australia. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources (pp. 181-186). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22017.html

[51] sign-lang@LREC: Workshop Series on the Representation and Processing of Sign Languages at LREC (International Conference on Language Resources and Evaluation). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/index.html

[52] Berez-Kroeker A.L., McDonnell, B., Koller, E. and Collister, L.B. 2022. *The Open Handbook of Linguistic Data Management*. MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001

[53] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J. et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data 3,* 160018. https://doi.org/10.1038/sdata.2016.18

©2021-2023 EASIER Consortium

Funded by the Horizon 2020 Framework Programme of the European Union