

How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages

Marc Schulder, Thomas Hanke

Institute of German Sign Language and Communication of the Deaf

University of Hamburg, Germany

marc.schulder@uni-hamburg.de, thomas.hanke@uni-hamburg.de

Abstract

The publication of resources for minority languages requires a balance between making data open and accessible and respecting the rights and needs of its language community. The FAIR principles were introduced as a guide to good open data practices and they have since been complemented by the CARE principles for indigenous data governance. This article describes how the DGS Corpus implemented these principles and how the two sets of principles affected each other. The DGS Corpus is a large collection of recordings of members of the deaf community in Germany communicating in their primary language, German Sign Language (DGS); it was created to be both as a resource for linguistic research and as a record of the life experiences of deaf people in Germany. The corpus was designed with CARE in mind to respect and empower the language community and FAIR data publishing was used to enhance its usefulness as a scientific resource.

Keywords: FAIR principles, CARE principles, Linguistic Corpus, German Sign Language

1. Introduction

When creating resources relating to minority groups the aims of open science and open data must be balanced against the needs and rights of the minority group. The FAIR principles (Wilkinson et al., 2016) are designed as a guide to good open data practices, but do not take into account the needs of minority groups. The CARE principles (RDA International Indigenous Data Sovereignty IG, 2019) have been introduced as a complementary guide to ensuring that data is not only open, but also respects indigenous and minority group stakeholders. In many aspects CARE principles can be used to guide how FAIRness can be implemented. In other aspects, CARE helps identify how openness of data should be limited or adjusted to protect and empower the language community.

In this article we discuss how the CARE and FAIR principles have affected the creation and publication of the *DGS Corpus*, one of the largest signed language corpora available as of early 2022. The *DGS Corpus* is a collection of recordings of German Sign Language (DGS; *Deutsche Gebärdensprache*) as used by members of its language community.

The primary stakeholders in the DGS language community are members of the deaf community in Germany. Their life experience differs distinctly from that of the majority population in a variety of factors, such as language barriers, accessibility of public services, medical treatment and education, personal identity and deaf-centric aspects of social life. As such they are a minority group in both a cultural and linguistic context. Consequently, CARE principles had to be a central tenet of the *DGS Corpus* creation process to ensure that it would contribute to the representation and empowerment of its community and avoid harm and exploitative practices.

FAIR, like CARE, represents a set of guiding principles, not a fixed set of steps and technologies to use. As such, strategies had to be developed for how to best publish a linguistic dataset as large and complex as the *DGS Corpus* in a FAIR manner so that it would actually improve it. These strategies cover technical decisions, such as how to allow the persistent referencing of individual parts of the data, but also how FAIR can actively be used to support CARE, e. g. by providing open data in ways that benefit the community.

The remainder of this article is structured as follows: [Section 2](#) describes the basic components of the FAIR and CARE principles and [Section 3](#) provides a general introduction to the *DGS Corpus*. [Section 4](#) then describes how various decisions during the creation and publication of the corpus were influenced by the considerations underlying CARE. These form the basis for how FAIRness in the resulting datasets can be implemented, which is described in [Section 5](#). [Section 6](#) provides a concluding summary and outlines potential future steps for improving the CARE and FAIRness of the corpus further.

2. Background

2.1. The FAIR Principles

The FAIR guiding principles (Wilkinson et al., 2016) stipulate that good data should be

Findable Data should be easy to find for both humans and machines. This requires globally unique and persistent identifiers which are indexed in searchable resources and associated with rich metadata.

Accessible Users need to know how to access (meta)data, possibly including steps for authentication and authorisation. Access should be de-

fined by metadata and use free and open protocols. Even when data is no longer available, its metadata should be.

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.

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.

Each of the four FAIR principles is further divided into a number of aspects that contribute to it. To fulfil the FAIR principles, data should be implemented in ways that make it both human- and machine-readable. For a more detailed description of the FAIR principles, see [Wilkinson et al. \(2016\)](#) and the *GO FAIR* website¹.

While they are related to the open data movement, the FAIR principles acknowledge that there are legitimate reasons to restrict access to some kinds of data, so data can be FAIR without having to be open ([Mons et al., 2017](#)).

2.2. The CARE Principles

While open data and FAIR focus on increasing the sharing of data, they do not address how such practices can be ethically implemented when the data in question originates within minority groups.

The *CARE Principles for Indigenous Data Governance* ([RDA International Indigenous Data Sovereignty IG, 2019](#)) were introduced to fill this gap, working as a response and complement to FAIR and as a counterbalance to open data requirements.

The CARE principles stipulate that the following should be observed:

Collective Benefit Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.

Authority to Control Indigenous Peoples' rights and interests in Indigenous data must be recognised and their authority to control such data be empowered.

Responsibility Those working with Indigenous data have a responsibility to share how this data is used to support Indigenous Peoples' selfdetermination and collective benefit.

Ethics Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

¹<https://www.go-fair.org>

As with FAIR, each CARE principle is subdivided further into concrete aspects. For detailed information on these, see the *CARE website*.²

While the CARE principles were designed with a focus on indigenous populations, they are also largely applicable to minority populations such as deaf communities ([Batterbury et al., 2007](#); [Bone et al., 2021](#)).

3. The DGS Corpus

The *DGS Corpus* is a part of the *DGS-Korpus project*, which is a long-term project to create both a corpus and dictionary of German Sign Language ([Prillwitz et al., 2008](#)). It was started in 2009 and has a runtime of 15 years. Its goals are the creation of a reference corpus (the *DGS Corpus*), the release of a public subset of this corpus with high quality annotations (the *Public DGS Corpus*) and the creation of a corpus-based dictionary (*DW-DGS — Digital Dictionary of German Sign Language*). This article focuses on the first two goals and does not further address the corpus-based dictionary, which is at a preliminary stage at the time of writing.

3.1. The Reference Corpus

The *DGS Corpus* is a reference corpus that consists of 560 hours of conversations in DGS. It involves 330 participants from all parts of Germany. They were grouped in pairs and provided with a number of conversational tasks, such as discussing a given topic or historical event, free dialogue or the retelling of stories ([Nishio et al., 2010](#)). Recordings were made between 2010 and 2012 and everyone present during recording sessions (including moderators and technical personnel) used DGS as their primary language. For more information on the curation and demographic of the corpus, see [Schulder et al. \(2021\)](#).

To open up the recordings to linguistic research, gloss annotations³ and translations into German and English are being made. Creating annotations for signed languages is an elaborate process that requires considerable work. While significant parts of the corpus have already been transcribed, the work is still ongoing and is not expected to be fully concluded by the end of the project.

Access to the full reference corpus is restricted, requiring a licence agreement that is contingent on an evaluation of the intended use case (see [Section 4.5](#)). However, part of the corpus is made publicly available. This public component is described in the following section.

²<https://www.gida-global.org/care>

³Signed languages usually have no written form, so their linguistic annotation relies on glosses. These usually consist of a spoken language word that represents an approximate lexical translation, an index to disambiguate distinct signs that received the same gloss word, and sometimes additional information. Glosses are metalinguistic labels, they do not encode all nuances of an utterance and are not context-appropriate translations.

3.2. The Public Corpus

The *Public DGS Corpus* is a publicly released subset of the reference corpus. It covers 50 hours of recordings and is fully annotated and translated. Its contents have been selected to both be interesting to a general audience and to give an impression of the different elicitation tasks that comprise the reference corpus. The annotations and translations have undergone additional quality assurance steps (Konrad et al., 2020) and personal information of participants and third parties has been anonymised (Bleicken et al., 2016).

The *Public DGS Corpus* was first released in 2018 (Jahn et al., 2018) and has since been updated with additional content and features about once a year (Hanke et al., 2020). It is published in two forms, *My DGS*⁴ and *My DGS – annotated*⁵, which differ in what information is provided and how it is presented. Each can be reached through its own website.

My DGS (Hanke et al., 2018) is a community portal for the deaf community, DGS teachers and others interested in DGS. Its videos can be viewed online and are searchable by topic, conversation format, region and participant age group. Optional German subtitles are available.

My DGS – annotated (Konrad et al., 2018), a research portal for the international scientific community. It provides a fully annotated corpus of DGS including translations, body pose data for automated processing, and metadata. Its data can be downloaded or displayed in an online annotation viewer. The web portal, annotations and translations are all available in both English and German.

Apart from providing access to the transcripts of individual recordings, the research portal also includes a **type index**.⁶ The index lists the type glosses for all signs encountered in the corpus, including how often each one occurs. For each type it leads to a page that lists all their token occurrences, showing their immediate context (associated utterance translation and neighbouring glosses). The page also specifies the citation form of the sign via a studio recording (where available) and phonetic transcription using HamNoSys (Hanke, 2004), and provides links to other lexical resources that describe the same sign (Müller et al., 2020).

4. CARE in the DGS Corpus

Although the CARE principles themselves were only published in 2019, ten years after the start of this corpus project, the underlying discourse on research ethics that would eventually result in it was already present,

⁴<http://meine-dgs.de>

⁵<http://ling.meine-dgs.de>

⁶As is customary in sign language corpora, the annotation distinguishes between sign types (the base citation form as it would be found in a dictionary) and sign tokens (the specific realisation in an utterance). Both are represented by glosses.

regarding both indigenous and deaf populations (Harris et al., 2009; Linguistic Society of America, 2009).

While the FAIR principles regard the publication of data and consequently mostly affect the final stages of the data creation cycle, CARE affects all stages of data creation and must be taken into account from the very start. In this section we outline how CARE affected the data collection phase (Section 4.1), the handling of informed consent (Section 4.2), in what forms data is presented to the public (Section 4.3), considerations on personal privacy and authorship (Section 4.4), public access and licensing (Section 4.5), and what amount of data could be released as part of the public corpus (Section 4.6).

For further details on the design and curation of both reference and public corpus, see also Schulder et al. (2021).

4.1. Data Collection

The DGS-Korpus project was designed from the ground up with the needs and rights of the German deaf community (the primary stakeholders within the DGS language community) in mind. Following the principle “nothing about us without us”, the project has always included several deaf team members. In addition, a focus group of deaf users was formed to guide project decisions. Project and focus group members also regularly participate in deaf-centric events to inform the community about the progress of the project and collect feedback.

To ensure *collective benefit*, the corpus was designed so that its content would function both as a source for linguistic research and as a record of deaf culture. Through a variety of elicitation tasks, participants were encouraged to share their general life experience, deaf-specific experiences (e.g. schooling and deaf clubs), their perception of specific historical events, tell jokes, etc. (Nishio et al., 2010). Tasks and topics were chosen to be of interest to the deaf communities, so that the resulting resource would be entertaining, informative and support the identity of the community. (Blanck et al., 2010)

4.2. Informed Consent

As a basis to *authority to control*, informed consent was requested from all corpus participants (Hanke et al., 2010). The consent process involved information about the purpose of the project, the collection, use and sharing of data, and the participant’s rights. All information was provided in both DGS (video) and German (text). Participants received the opportunity to ask clarifying questions. Consent was then received as a recorded utterance in DGS and via signature.

The provided consent allows the use of the recorded data by the project. Regarding the sharing of data with third parties, participants could choose to opt out, choose on a case by case basis or delegate the decision to the project. This decision could be made separately for the sharing of data for non-commercial purposes of

(a) teaching or (b) cultural heritage and (c) the sharing of contact data with other researchers. Based on their decisions, three participants were excluded from the public part of the corpus, while the remaining 327 are all represented in it.

After recordings were concluded, participants were sent copies to review. They were given the opportunity to give or withhold approval and to identify specific sections that they would like excluded. This resulted in 60 exclusions of a few seconds or minutes, totalling 48 minutes – only 0.1 percent of recorded material – and no general exclusions. For details, see [Hanke et al. \(2010\)](#).

Furthermore, participants were informed that they retain permanent legal control over their recordings and are free to request the removal of recordings at any time. As of the time of writing, no requests have been made.

4.3. Data Presentation

As mentioned in [Section 4.1](#), the DGS corpus is both a linguistic dataset and a record of deaf culture. To ensure that both these goals would be met by the public corpus and maximise the *collective benefit* it provided, it was decided to create two independent portals (see [Section 3.2](#)), each optimised for one of its two goals.

The portal *My DGS* was designed with the deaf community as its primary user group in mind and DGS teachers and people generally interested in DGS and deaf culture as secondary user groups ([Jahn et al., 2018](#)). This involved focusing navigation on user interests, such as finding conversations about specific topics or by geographic region.

As DGS was the primary user language, German was used only where textual elements could not be avoided. The only longer description, an introduction to the project, was provided in both DGS and German. For the recordings, optional German subtitles are provided. The exception are recordings in which participants were asked to tell jokes. These are not subtitled, as spoken language translations were found to not capture their humour sufficiently.

Some data was omitted as it was considered to not be interesting to a non-scientific audience, such as the gloss annotations and certain recordings that covered research-focused tasks, such as the retelling of picture stories commonly used in linguistic data collections.

A preliminary version of the community portal with limited data was released in late 2015 to receive feedback from the deaf community and improve its features for the first release of the full dataset in 2018. Further feedback since then has resulted in additional changes. For example, the interface of the research portal, *My DGS – annotated*, was originally only made available in English, as its target audience was the international research community. Following interest from the general deaf community to inspect the research aspects of the corpus as well, a German interface was added.

4.4. Privacy and authorship

In Germany, personal privacy is considered an important right. While the nature of the DGS corpus entails that participants can not have full anonymity as they are shown on video and asked to talk about personal experiences, the project deemed it important to protect their privacy and that of third parties as far as possible in the interest of *minimising harm*.

In all recordings, personally identifiable information (such as names, dates of birth or places of residence with small populations) of anyone not considered a public figure in the general or deaf population was removed. This anonymisation process affects videos, annotations, translations and pose information. ([Bleicken et al., 2016](#); [Isard, 2020](#)).

Both the reference and public corpora are designed to provide a diverse and balanced selection of signers from different regions, ages and genders. In the public corpus, all recordings provide metadata to identify these factors, but to protect participant privacy, region and age are simplified to broad categories. Other recorded metadata, such as educational background and age of language acquisition, are not made public.

Another requirement of CARE is to establish provenance, i.e. where data originated. While the privacy considerations of the public corpus obfuscate that information, it can be clearly reconstructed through the restricted-access records of the reference corpus, in which each recording is associated with the identity and contact information of its participants. This way, authorship can be established at any time.

4.5. Access and Licensing

When choosing a licence for the Public DGS Corpus, CARE responsibilities had a strong impact on how open data conditions could be achieved. While commonly used open licences, such as those by Creative Commons, would allow open access and thus fulfil FAIR requirements, they were deemed too permissive to ensure the *responsible* and *ethical* use of the public corpus. Instead, custom licence conditions were developed.

Each portal received a separate licence that reflects its intended uses. The community portal *My DGS* may be viewed for private and non-commercial purposes. The downloading of contents and integration into other services is limited to teaching purposes. Data from the research portal *My DGS – annotated* may only be used for linguistic research.

To prevent the exploitation of its participants, commercial use of the corpus is not permitted. The limitation of research uses to linguistics was chosen to protect against data being used in contexts not related to DGS and deaf culture, such as general purpose image recognition, or in applications that might pose harm to individuals or the community.

To use the public corpus for applications not covered by its public licences (e. g. for other research area) or to ac-

cess restricted parts of the reference corpus, a separate and explicit licence agreement is required. Such agreements are negotiated on a per-case basis and contingent on evaluation of their *collective benefit* and *ethics*. If use case is not covered by the consent provided by participants, additional consent has to be requested.

4.6. Size of Public Corpus

For a recording to be publishable in the public corpus, its translation and annotation must be complete and of sufficient quality. Personally identifiable information must be anonymised (see [Section 4.4](#)). Processing one hour of corpus recording to meet these standards requires about 800 to 1000 work hours. This puts hard constraints on how much of the reference corpus could be included in the public corpus.

Quality assurance procedures for translation and annotation are interconnected. Translations of all corpus recordings were created early in the project to function as an additional aid to annotators. At the same time, the annotation process is used to highlight mistranslations. This is more efficient than running an independent verification of all translations. As flawed translations can potentially reflect badly on a participant by misrepresenting their views, quality assurance is an important part of CARE *responsibility*.

Any recording that touches upon the private lives of the participants must be sufficiently anonymised. As the DGS Corpus has a strong focus on open conversations and the discussion of life experiences, this affects it more strongly than it does corpora which focus on linguist tasks that do not touch on the personal lives of participants. To identify which utterances require anonymisation, the project relies on translations and annotations (see [Bleicken et al. \(2016\)](#)).

From these constraints follows a hierarchy of minimum work load for the publication of different parts of the corpus. For any task involving personal experiences (the majority of recordings) full annotation, translation and anonymisation is required. Anonymisation can be skipped for tasks revolving around entirely fictional contents, such as jokes and the retelling of narratives that were provided by the project. In the case of jokes it was also determined that they could still provide value without annotation and translation.

For narrative retellings and similar research-oriented tasks, the decision whether to invest in publishable annotations and translation was less straightforward. Originally it was decided that such recordings would not be useful enough to researchers without annotation, so only fully annotated recordings were included. This decision has been revisited recently, in part because the automatically generated pose information that was added to the corpus in releases 2 and 3 may in some cases work as a stand-in for missing annotations. Accordingly, additional recordings without annotation or translation will be published in the upcoming fourth release of the public corpus.

5. A FAIR DGS Corpus

Releasing a subset of the DGS Corpus publicly was part of the project plan from the beginning ([Prillwitz et al., 2008](#)). The expectation of what exactly such a public release should entail developed over the years as best practices in scientific data publishing evolved. An important milestone in this evolution were the FAIR principles introduced in 2016.

The FAIR principles are guidelines that describe what conditions should be met to create good scientific datasets, but they do not specify how exactly these conditions should be fulfilled. This is by design, as appropriate solutions depend on the nature of the dataset, the best practices of a scientific field and the technical solutions available at the given time.

This section describes how the DGS Corpus implements the FAIR principles. As a large linguistic corpus published as two separate but related datasets, the public corpus requires a complex structure of interconnected persistent identifiers ([Section 5.1](#)). Extensive metadata is provided, both relating to the description of data in general and linguistic data in particular ([Section 5.2](#)). The original full recordings of the reference corpus, while restricted access, are also equipped with identifiers and metadata and put in relation with the data of the public corpus ([Section 5.3](#)). Finally, the documentation of the reference and public corpus is published in a FAIR manner to support the data ([Section 5.4](#)).

5.1. Persistent Identifiers for the Public Corpus

The most fundamental requirement of FAIR is to provide a dataset with a persistent identifier to ensure it is always *findable* in the same way. For small static datasets this can be a single identifier. For larger and more complex datasets that get updated over time, it is advisable to also provide a set of interlinked identifiers. This allows users to specify exactly which part and version of a dataset they are referencing.

The *DGS Corpus* uses Digital Object Identifiers (DOIs), one of the most commonly used persistent identifiers for digital data. DOIs are centrally registered and can be provided in the form of a URL. Following such a DOI URL then forwards the user to the current location of the dataset. This is more persistent than providing the direct URL of the current location, as it can easily be updated in case the location (and therefore the direct URL) changes. In addition, it can be used to explicitly associate metadata with the object that the DOI represents (see [Section 5.2](#)).

Each DOI in the *Public DGS Corpus* is used to uniquely identify a specific component. This can be the whole dataset ([Section 5.1.1](#)), a specific recording transcript ([Section 5.1.2](#)) or a particular sign type ([Section 5.1.3](#)). To differentiate between different releases of the same component, a set of versioned DOIs is registered for each component ([Section 5.1.4](#)).

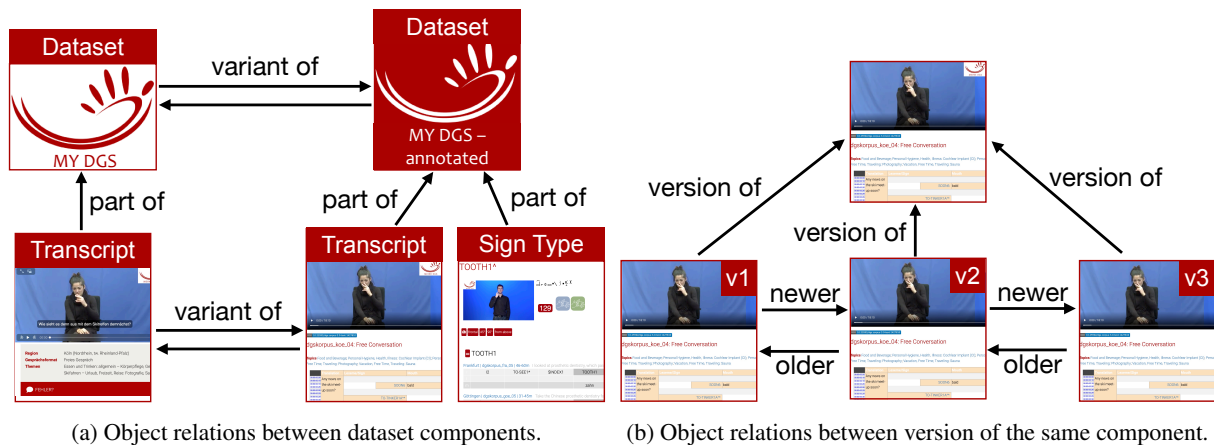


Figure 1: Visual representation of DOI object relations in the *Public DGS Corpus*.

5.1.1. Dataset

As mentioned in Section 3.2 the Public DGS Corpus is published as in two forms: *Meine DGS* and *Meine DGS – annotated*. These are treated as distinct datasets that are related but distinct from one another. As such, each dataset receives its own set of DOIs. This involves DOIs representing the dataset as a whole as well as DOIs representing individual components (described in Sections 5.1.2 and 5.1.3).

5.1.2. Transcript

The public corpus is separated into individual transcripts. Each transcript represents a single elicitation task from a specific recording session between two participants. The transcript consists of the video recording, its annotation and translations, and other associated data, such as pose information and metadata files. Each transcript has its own set of DOIs to allow users to reference particular discourses.

5.1.3. Sign Type

The type index of the research portal (see Section 3.2) provides an alternative way of viewing the dataset from the perspective of individual signs and their distribution across the dataset. Each sign type entry receives its own set of DOIs.

This is the most unusual kind of DOI in the corpus, as it references a conceptual entity rather than one that follows from the primary data structure.

Sign type DOIs are included to support researchers who wish to refer to specific signs in their research. This can for example be useful for corpus analyses focusing on particular signs, cases where it would be impractical to reference all relevant transcript locations in which a sign occurs, or for researchers that use the type index as a lexical resource, referring to the sign in general, rather than to a specific instance in the corpus.

Identifying sign types by DOI also helps provide persistence when the gloss of a sign type changes. While this does not happen often, it may occasionally be the case when the spoken language part of a gloss is found to be erroneous or insufficient in some way. In such

cases, the type DOI will still correctly refer to the same sign, as it represents the sign language sign, not the spoken language gloss.

5.1.4. Version

Since its original release in 2018 the public corpus has been regularly updated and extended (Hanke et al., 2020). The data of each individual release is frozen and does not change anymore. While the main URLs of the portals always refer to the latest release, old releases are still accessible.

New DOIs are created for each release of the corpus to allow users to clearly specify which release they are referring to, e. g. to ensure the replicability of their research.

Each DOI is intended as a unique identifier of an object, so only if the object in question has changed between versions is a new DOI assigned, otherwise the object and its DOI are understood to be equally related to multiple corpus releases. A transcript receives a new DOI if at least one of its files is changed, for example due to corrections in the annotation or updates to its pose information. Similarly, sign types receive a new DOI if their set of token instances changes (due to corrections or new data) or other information such as the gloss or HamNoSys transcription is updated. In the case of DOIs for the complete *My DGS* and *My DGS – annotated* datasets, a new DOI is assigned for each release, as a new release will by definition include changes to the dataset.

In addition to the release-specific DOIs described so far, so called *Version DOIs*, all corpus objects also receive a *Concept DOI*. This is a DOI that refers to the object without specifying which version is meant. While for most cases a Version DOI should be used to be as specific as possible, in some cases it is preferable to clearly state that the object is referred to in general or that the most recent available version should be used.

5.2. Metadata for the Public Corpus

Metadata is provided through two mechanisms: DOI metadata for general purpose dataset information (Sec-



Öffentliches DGS-Korpus Release 3.0 / Public DGS Corpus Release 3.0

 10.25592/dgs.corpus-3.0


Dieses Release / This release:

Öffentliches DGS-Korpus Release 3.0:	Webseiten DE
Public DGS Corpus Release 3.0:	Website EN

Ältere Versionen / Older versions:

Öffentliches DGS-Korpus Release 2.0:	 10.25592/dgs.corpus-2.0
Public DGS Corpus Release 2.0:	2.0
Öffentliches DGS-Korpus Release 1.0:	 10.25592/dgs.corpus-1.0
Public DGS Corpus Release 1.0:	1.0

Versionsunabhängige DOI / Version-independent DOI:

Jeweils neuestes Release des Öffentlichen DGS-Korpus:	 10.25592/dgs.corpus
Always the newest release of the Public DGS Corpus:	

Zitiervorschlag / Cite as:

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., Schulder, M. 2020. *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release* [Dataset]. Universität Hamburg. <https://doi.org/10.25592/dgs.corpus-3.0>

```
#misc{dgs_corpus_3,
  title = {MEINE DGS -- annotiert. (\^o)ffentliches Korpus der Deutschen Geb(\^a)rdensprache, 3. Release / MY DGS -- annotated. Public Corpus of German Sign Language, 3rd release },
  author = {Konrad, Reiner and Hanke, Thomas and Langer, Gabriele and Blanck, Dolly and Bleicken, Julian and Hofmann, Ilona and Jeziorski, Olga and K(\^o)nig, Lutz and K(\^o)nig, Susanne and Nishio, Rie and Regen, Anja and Salden, Uta and Wagner, Sven and Worseck, Sat u and R(\^a)lse, Oliver and Jahn, Elena and Schulder, Marri...
```

(a) Landing page for release 3 of the *My DGS – annotated* dataset. Provides links to the online portal of the corpus release as well as links to other releases. This is followed by information on how to cite the dataset.

dgskorpus_koe_05 – Erlebnisbericht / Experience Report

 10.25592/dgs.corpus-3.0-text-1428038

Diese Version / This version:

Im Öffentlichen DGS-Korpus Release 3.0:	Webseite DE
In the Public DGS Corpus Release 3.0:	Web Page EN

iLex v. 3.0	1428038.iLex
ELAN v. 3.0	1428038.eaf
Video A1 v. 1.0	1428038_1a1.mp4
Video B1 v. 1.0	1428038_1b1.mp4
Video C v. 1.0	1428038_1c.mp4
SRT v. 3.0	1428038_de.srt 1428038_en.srt
Video AB v. 1.0	1428038.mp4
OpenPose v. 3.0	1428038_openpose.json.gz
Metadata (CMDI) v. 3.0	1428038.cmdi

Versionsunabhängige DOI / Version-independent DOI:

Jeweils neueste Version im Öffentlichen DGS-Korpus:	 10.25592/dgs.corpus-text-1428038
Always the newest version in the Public DGS Corpus:	1428038

(b) Landing page for a specific transcript of release 3. Version information for each file indicates the release in which it was last changed.

Figure 2: Examples of DOI landing pages for the *My DGS – annotated* dataset of the public corpus.

tion 5.2.1) and metadata in CMDI format for information specific to language datasets (Section 5.2.2).

5.2.1. DOI Metadata

Machine-readable metadata is associated with every DOI. The DOI metadata of the public corpus follows the DataCite Schema (DataCite Metadata Working Group, 2019), which is designed to provide general information on research datasets.

Each DOI object in the corpus provides a title, description and set of keywords in English and German, identifies the authors, publisher, funding body, publication date, and file formats.

Each object also specifies how it is connected to other parts of the corpus through specific *related identifier* relations. These relations indicate, where applicable, which dataset the object belongs to, its older and newer versions, its version-agnostic concept entry and whether there is a corresponding object in the other corpus dataset. A visual representation of these relations can be seen in Figure 1.

Following best practices established for DOI redirects, users are sent to a dedicated DOI landing page, rather than to the data itself. This landing page provides the most relevant parts of metadata in a human-readable

form. Examples of such landing pages can be seen in Figure 2.

In all cases the landing page provides links to the data, other versions and the concept DOI. The remaining choice of metadata depends on the corpus object in question. Dataset objects provide their suggested citation format (see Figure 2a). Transcript objects also offer a list of the individual files belonging to the transcript (see Figure 2b). Each of these files has a direct download link and an indicator of which release the file was added or last changed for. For example, a release 3 transcript may have video files that were unchanged since the first release, pose data that was newly added in release 2 and annotation files that were changed for release 3 due to corrections to the annotation (thus resulting in the creation of a distinct DOI for this transcript release).

5.2.2. CMDI Metadata

The metadata provided with DOIs focuses on information that is generally relevant for the handling of a dataset. To also encode information specific to a linguistic corpus, every transcript includes a CMDI metadata file. The CMDI XML schema (ISO 24622-2:2019, 2019) was designed specifically to model information

on language datasets, such as metadata on participants, elicitation tasks, language variant, geographic location and the type of linguistic resource.

To improve findability and provide users with a more human-friendly interface to the CMDI data, the research dataset is indexed in the Virtual Language Observatory⁷ (Van Uytvanck et al., 2012).

5.3. Original Recordings

While the FAIR principles encourage open data, they are also of relevance to restricted data, particularly when it comes to data archival.

All original recordings of the DGS Corpus are long-term archived in the research data repository of Universität Hamburg to provide redundant long-term storage independent of the daily operations of the project.

Each recording session is assigned a DOI and provided with relevant metadata. While access to the actual data is restricted, all metadata is public. For recordings that were the basis of transcripts in the public corpus, this connection is indicated by appropriate *related identifier* relations (see Section 5.2.1).

5.4. Documentation

Apart from its metadata describing individual components of the corpus, the project has also released a lot of public documentation. In addition to peer-reviewed publications, 26 project notes have so far been published. These cover a variety of aspects, such as data collection, transcription, corpus publication and associated tools. They also include a *data statement* (Schulder et al., 2021), a document type specifically designed to aid users and developers in understanding the provenance and creation process of a dataset to judge its inherent biases (Bender and Friedman, 2018).

With respect to CARE, the documentation provides accountability, showing what measures were taken to create a beneficial resource and support others in using it appropriately. For FAIR they support *reusability* in general and *provenance* in particular.

Regarding the FAIR handling of the documentation itself, each project note is archived in the research data repository of Universität Hamburg, including DOIs, associated metadata and an open Creative Commons Attribution license. When documentation is updated, old versions remain available under their respective version DOI. The collection in its entirety is also identified by a DOI.⁸

6. Conclusion

In this article we presented how the *DGS Corpus*, a large corpus of German Sign Language, was created following both CARE and FAIR principles. As a linguistic dataset for a cultural and linguistic minority, CARE guided the design of the corpus and informed how FAIR publishing principles might be implemented

in an ethical manner. Both the open and restricted access parts of the corpus were influenced and enhanced by both FAIR and CARE.

The *DGS Corpus* also functions as an example of how FAIR principles can be applied to a complex linguistic dataset. This involves a concept for assigning persistent identifiers in ways that aid linguistic research practices, the creation and integration of diverse metadata, extensive open documentation and long-term archival strategies.

The CARE and FAIRness of the *DGS Corpus* have been extended step by step over time. We plan to continue this work in future corpus releases. Possible improvements include extending DOI functionality to allow the referencing of exact timestamps in transcripts, publishing additional recordings and fully interconnecting its DOI relation structure with the upcoming digital dictionary of DGS.

7. Acknowledgements

The authors would like to thank Amy Isard for her valuable feedback.

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

8. Bibliographical References

- Batterbury, S. C. E., Ladd, P., and Gulliver, M. (2007). *Sign language peoples as indigenous minorities: Implications for research and policy*. *Environment and Planning A: Economy and Space*, 39(12):2899–2915. DOI: [10.1068/a388](https://doi.org/10.1068/a388).
- Bender, E. M. and Friedman, B. (2018). *Data statements for natural language processing: Toward mitigating system bias and enabling better science*. *Transactions of the Association for Computational Linguistics*, 6:587–604. DOI: [10/gft5d7](https://doi.org/10/gft5d7).
- Blanck, D., Hofmann, I., Jeziorski, O., König, S., Langer, G., and Rathmann, C. (2010). *Uses of the DGS Corpus from a deaf community perspective*. Poster presented at the 4th Workshop of the Sign Language Corpus Network, Berlin, Germany. DOI: [10.25592/uhhfdm.8259](https://doi.org/10.25592/uhhfdm.8259).
- Bleicken, J., Hanke, T., Salden, U., and Wagner, S. (2016). *Using a language technology infrastructure for German in order to anonymize German Sign Language corpus data*. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3303–3306, Portorož, Slovenia. European Language Resources Association.
- Bone, T. A., Wilkinson, E., Ferndale, D., and Adams, R. (2021). *Indigenous and deaf people and the implications of ongoing practices*

⁷<https://vlo.clarin.eu/>

⁸<https://doi.org/10.25592/dgs.korpus.aps>

- of colonization: A comparison of Australia and Canada. *Humanity & Society*, pages 1–27. DOI: 10.1177/01605976211001575.
- DataCite Metadata Working Group. (2019). *Datacite metadata schema documentation for the publication and citation of research data v4.2*. DOI: 10.5438/BMJT-BX77.
- Hanke, T., Hong, S.-E., König, S., Langer, G., Nishio, R., and Rathmann, C. (2010). *Towards fair licences for data from the DGS Corpus project*. Poster presented at the 4th Workshop of the Sign Language Corpus Network, Berlin, Germany. DOI: 10.25592/uhhfdm.1885.
- Hanke, T., Schulder, M., Konrad, R., and Jahn, E. (2020). *Extending the Public DGS Corpus in size and depth*. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association.
- Hanke, T. (2004). *HamNoSys – representing sign language data in language resources and language processing contexts*. In *Proceedings of the 1st Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 1–6, Lisbon, Portugal. European Language Resources Association.
- Harris, R., Holmes, H. M., and Mertens, D. M. (2009). *Research ethics in sign language communities*. *Sign Language Studies*, 9(2):104–131. DOI: 10.1353/sls.0.0011.
- Isard, A. (2020). *Approaches to the anonymisation of sign language corpora*. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives*, pages 95–100, Marseille, France. European Language Resources Association.
- ISO 24622-2:2019. (2019). *Language resource management — component metadata infrastructure (CMDI) — part 2: Component metadata specification language*. Standard ISO 24622-2:2019, International Organization for Standardization, Geneva, Switzerland.
- Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). *Publishing DGS Corpus data: Different formats for different needs*. In *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 107–114, Miyazaki, Japan. European Language Resources Association.
- Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2020). *Public DGS Corpus: Annotation conventions*. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany. DOI: 10.25592/uhhfdm.1860.
- Linguistic Society of America. (2009). *Linguistic Society of America ethics statement*.
- Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., and Wilkinson, M. D. (2017). *Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud*. *Information Services & Use*, 37(1):49–56. DOI: 10.3233/ISU-170824.
- Müller, A., Hanke, T., Konrad, R., Langer, G., and Wähl, S. (2020). *From dictionary to corpus and back again – linking heterogeneous language resources for DGS*. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 157–164, Marseille, France. European Language Resources Association.
- Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). *Elicitation methods in the DGS (German Sign Language) Corpus project*. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 178–185, Valletta, Malta. European Language Resources Association.
- Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). *DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German*. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association.
- Research Data Alliance International Indigenous Data Sovereignty Interest Group. (2019). *CARE principles for indigenous data governance*.
- Schulder, M., Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., König, L., König, S., Konrad, R., Langer, G., Nishio, R., and Rathmann, C. (2021). *Data statement for the Public DGS Corpus*. Project Note AP06-2020-01, DGS-Korpus project, IDGS, Universität Hamburg. DOI: 10.25592/uhhfdm.9700.
- Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). *Semantic metadata mapping in practice: the Virtual Language Observatory*. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1029–1034, Istanbul, Turkey. European Language Resources Association.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., et al. (2016). *The FAIR guiding principles for scientific data management and stewardship*. *Scientific Data*, 3:160018. DOI: 10/bdd4.

9. Language Resource References

- Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2018). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache*. DGS-Korpus project, IDGS, Hamburg University, DOI: [10.25592/dgs.meinedgs](https://doi.org/10.25592/dgs.meinedgs).
- Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., and Worseck, S. (2018). *MY DGS – annotated. Public Corpus of German Sign Language*. DGS-Korpus project, IDGS, Hamburg University, DOI: [10.25592/dgs.corpus](https://doi.org/10.25592/dgs.corpus).