Persistent Identifiers and Metadata for the Public DGS Corpus

Author: Thomas Hanke

DOI (latest version): 10.25592/uhhfdm.10219

Releases:

v 0	2021-07-03	Initial draft
v 1	2021-12-30	First complete version, indicating changes to expect for Release 4 of the DGS Public Corpus

Abstract

We describe the granularity and scope of persistent identifiers assigned to both primary data (field data and studio recordings) collected in the DGS-Korpus project and the secondary data published as the Public DGS Corpus. In addition, we describe the corpus metadata that are made available as part of the Public DGS Corpus since its release 3.

Zusammenfassung (German Abstract)

Wir beschreiben die Granularität und persistenter Identifikatoren, die für Primärdaten (Felddaten und Studioaufnahmen) aus dem Projekt DGS-Korpus sowie für Sekundärdaten, die als Öffentliches DGS-Korpus veröffentlicht sind, vergeben wurden. Des Weiteren beschreiben wir die Korpus-Metadaten, die seit Release 3 als Teil des Öffentlichen Korpus verfügbar gemacht werden.

Introduction

The primary data collected in the DGS-Korpus project consists of two kinds:

- The DGS Corpus consists of 165 sessions of video recordings with multiple cameras, showing the two informants plus the moderator involved. Recording durations ranged from 255 to 300 minutes. The number of cameras was 7 for the recordings done in 2010 (two of them stereo cameras), and 8 thereafter (three of them stereo cameras). One recording session typically consists of about 20 tasks, i.e. data collection settings like discussions, story retellings etc, reporting personal experiences etc.
- Studio recordings are multi-camera recordings of varying length, typically to document sign types by one person signing them. In most cases, 4 cameras are involved (3 of them stereo) to show the signer from different perspectives (frontal, 45°, side, bird's eye).

The Public Corpus releases consist of approx. 50 hours of signing selected from the DGS Corpus. They are organised in sessions that are actually parts of the tasks mentioned above. Altogether, 327 participants in the recordings are represented in the Public Corpus. The videos contained in the Public Corpus are scaled down from the primary data, leaving out the bird's eye view camera and with blurring applied where necessary for anonymisation purposes.

Both the primary video data collected in the project and the Public DGS Corpus data have been assigned persistent identifiers. We describe the granularity these identifiers have been assigned with as well as the metadata provided in the identifier registration process. Metadata more specific to language corpora are part of the iLex annotation files as well as separate CMDI files (from release 3 of the Public DGS Corpus on). Here we describe our choices in mapping DGS-Korpus internal metadata to the CMDI profile selected.

Persistent Identifiers on Primary Data

The original recordings are archived in the University of Hamburg Research Data Repository session by session. Concept and version-specific DOIs are automatically assigned when submitting data. As the data is not supposed to change as long as the original recordings' file format (.mov) and codec (mostly Apple ProRes) can be read by video processing tools, most deposits have only one version. Metadata provided are restricted to what the archiving process requires. The participants are listed by their codes. The description text contains a table with the data necessary to provide start-to-start synchronisation of the video files within one session.
The data in the repository is meant as an archive to be accessed should the working copies of the movies held on a central file server get lost or if progress in movie encoding technology makes re-encoding from the originals necessary. As the datasets contain the original recordings without blurring applied where necessary, these repository datasets are "closed access".
Both the concept DOI and the version-specific DOI in the Research Data Repository have the format `10.25592/uhhfdm.N` with *N* being a counter. Due to the way the Datacite software works, the first version of the deposit gets *N+1* if *N* is the counter value of the concept DOI. For the field data, *N* is between 953 and 8955, newer studio recordings of course have higher numbers assigned.
Currently, there is no persistent identifier for the set of archival records.

Persistent Identifiers on Secondary Data

The Public DGS Corpus is presented on two separate portals, one mainly targeting the language community, the other one targeting the research community.
For the DOIs assigned, different name spaces are used for the two portals:

- `10.25592/dgs.meinedgs` for the language community portal
- `10.25592/dgs.corpus` for the research community portal

Actually, these two DOIs serve as the concept DOIs for the two portals, the version-specific DOIs are formed by appending the version number, e.g.

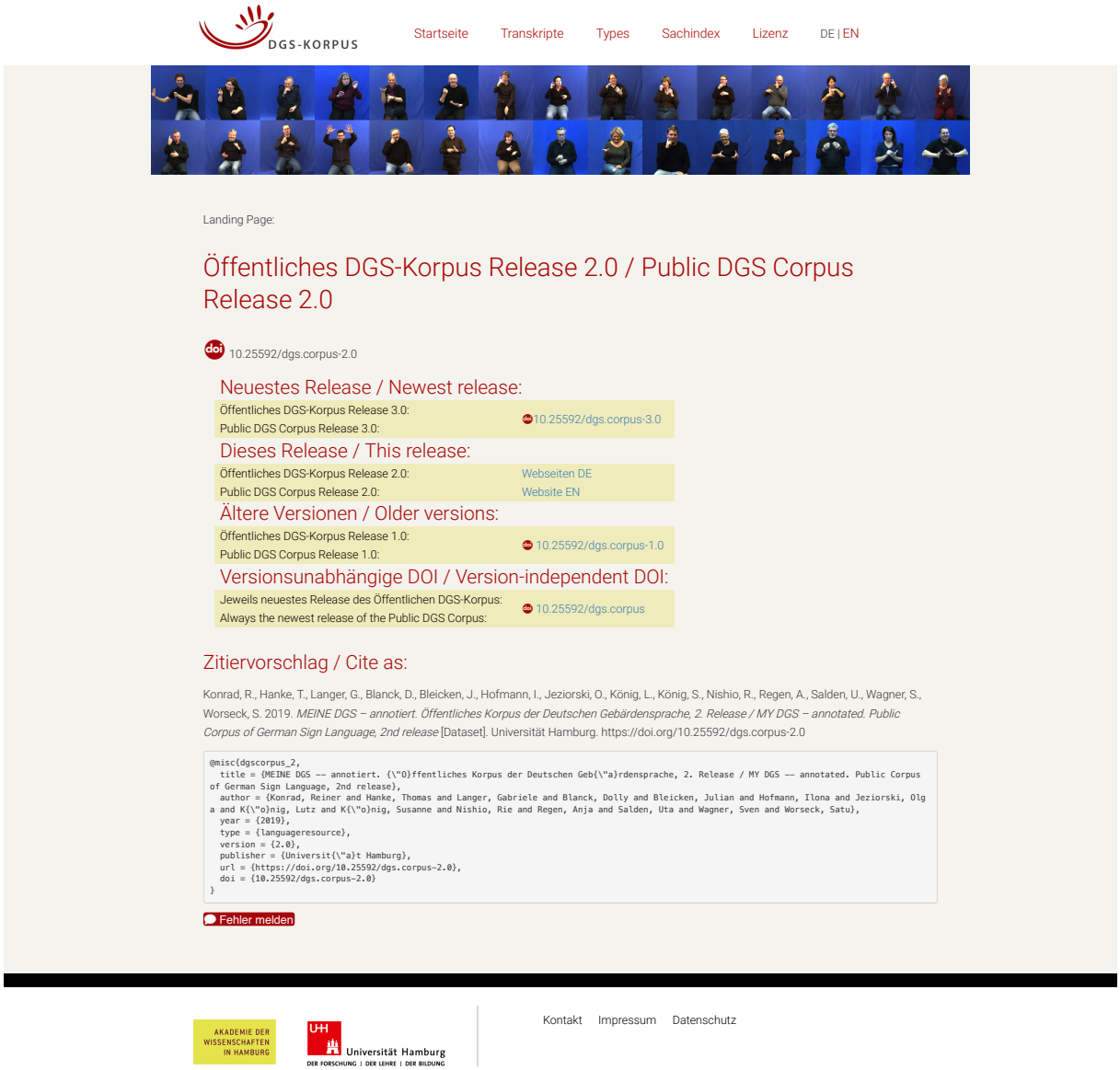
- `10.25592/dgs.corpus-3.0`

DOIs have also been assigned to individual sessions of the Public Corpus data. This leads to DOIs of the form

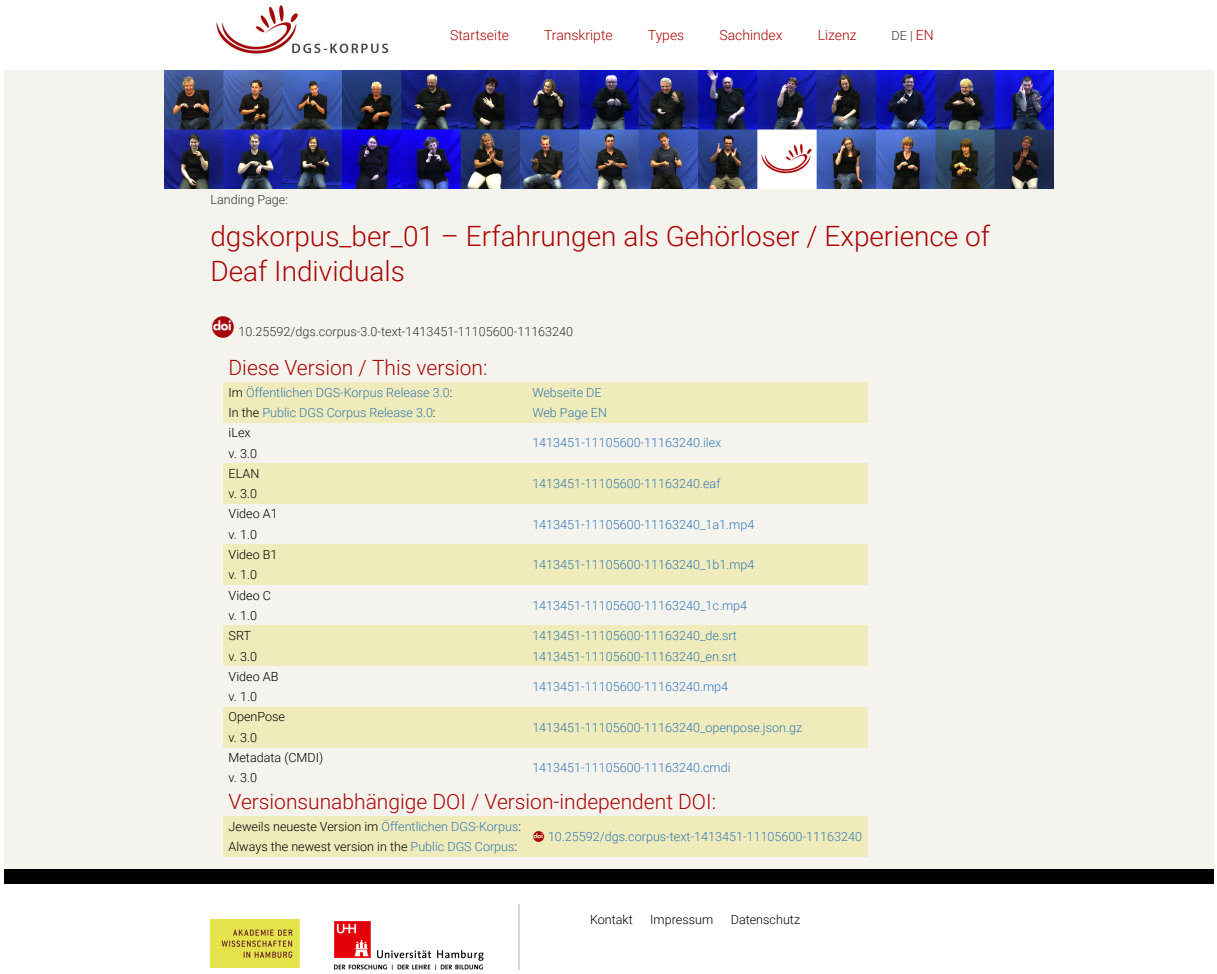
- `10.25592/dgs.corpus-3.0-text-1413451` (individual session in research portal belonging to version 3 of the Public Corpus)
- `10.25592/dgs.meinedgs-1.0-video-1250923` (individual session in language community portal belonging to version 1 of the Public Corpus)

The DOI of a session changes from version to version as soon as one file contained in the dataset differs between the releases.
In contrast to the DOIs pointing to the Research Data Repository, the URLs associated to the DOIs for the Public Corpus always lead to a landing page which gives access to older and

newer versions of the same dataset as well as the parent dataset where applicable.



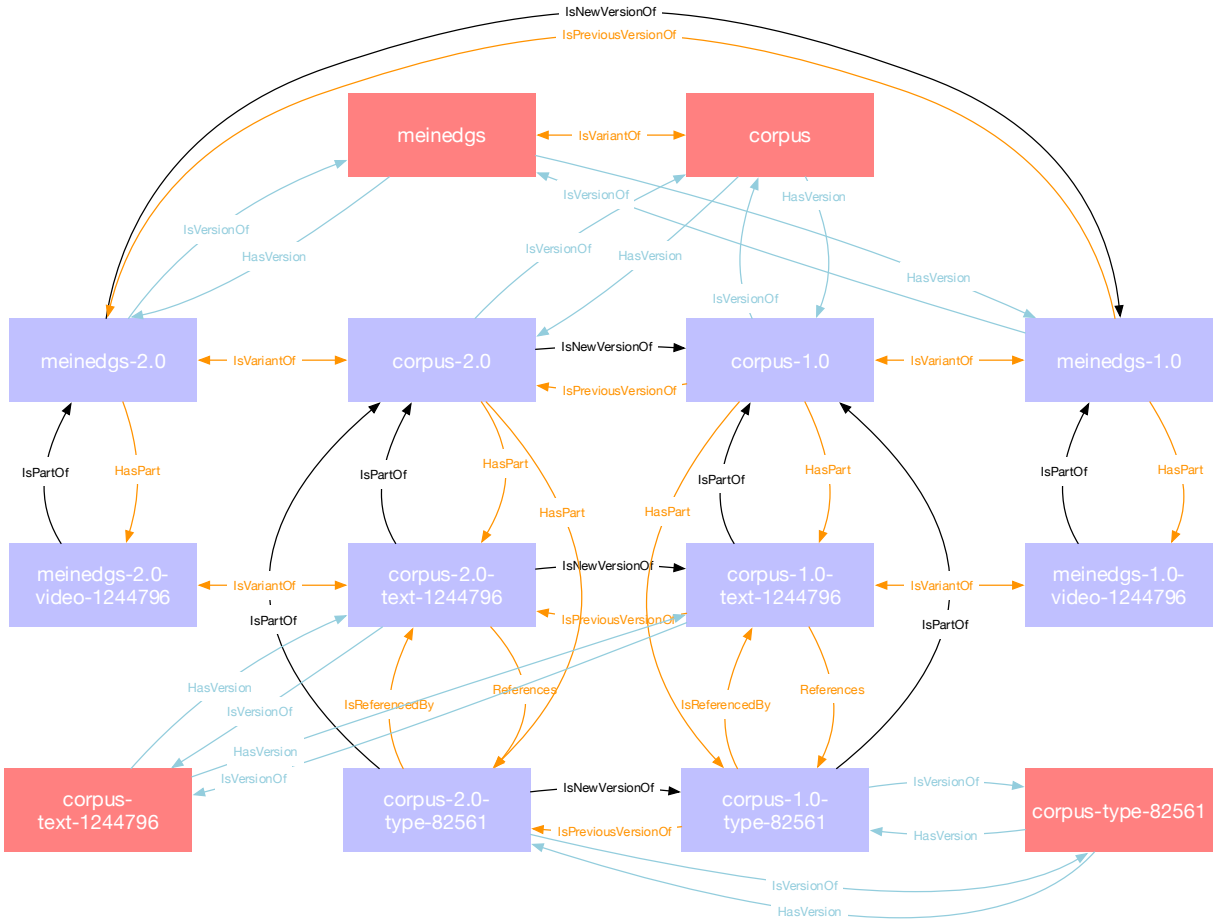
In the case of the research portal, the landing page also gives the choice between the German and English version of the website. The landing page also lists the individual components (files) that the dataset DOI points to.



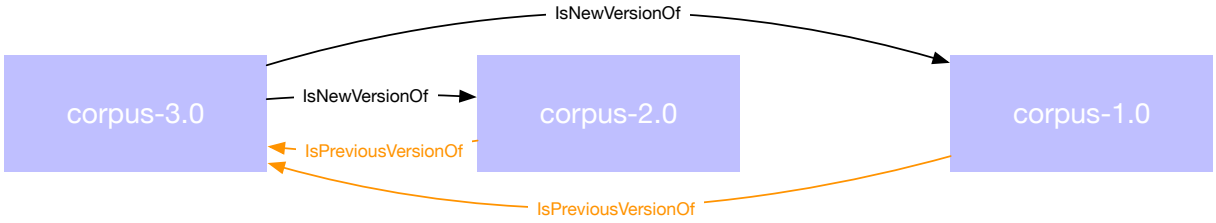
In addition, separate DOIs are provided for each of the type entries in the research portal from release 2 on. Release 4 will add DOIs for the elicitation task pages.

Relations between Persistent Identifiers within the Public Corpus

The DOI metadata provide more relation between the datasets of different granularity than the landing pages show. E.g. the corresponding pages from both portals are related by the attribute “is variant of”, the corpus DOIs list all their sessions under the attribute “has part”. The following figure gives the relations established the newest release entities and their immediate predecessors:



In order to restrict the relations to useful information, we do not provide a linkage between the releases. Instead, older versions only relate to the newest, and only the newest lists its predecessors:



Relations between Secondary and Primary Data

Starting with release 4 of the Public Corpus, deposit records for primary data will point to all Public Corpus sessions being annotated extracts from the recording session. These relations are tagged “has this upload as part“. This gives the user of the repository an idea of the closed-access content by viewing the public parts. In that respect, it does not matter that the Public Corpus records contain annotation that the primary data records do not contain. The reverse direction, i.e. from DOIs in the Public Corpus to the primary data archival records is not explicitly annotated as these links to not provide any useful extra information for the users.

Mapping of DGS-Korpus metadata to CMDI for the Public DGS Corpus

Introduction: Coverage and Aims

In releases 1 and 2 of the Public DGS Corpus, metadata was only made available via the web interface and through iLex files (which contain both metadata and transcription data), but was not directly accessible for ELAN or MaxQDA users. To close this gap, release 3 contains CMDI files (Component MetaData Infrastructure, https://www.clarin.eu/content/component-metadata) to provide the metadata we consider suitable to accompany the public data. These are:

Per participant:

Sex
Age range
Region
An identifier allowing users of the corpus data to determine all sessions involving that individual.

Per session (subtasks in project-internal parlance):

The participants involved.

In addition, the metadata may contain details on the recording sessions published elsewhere, including the elicitation setting used in the session, technical details of the recording, and the location where the recording took place.

So far, the exact dates of the recordings were not included in this list nor mentioned in publications. They are, however, part of the metadata of the archival copies of the original video files and as such become publicly available. Via the same participant identifiers used across the project, the recording dates for the sessions included in the Public Corpus can be inferred from there. As this information is not considered to touch personal rights, there is no reason not to include these dates in the metadata where the recording date is part of the metadata profile used.

Apart from making necessary meta-information available to people who have access to the session data, CMDI metadata are also used within the language resources community to include these resources in public catalogues such as the Virtual Language Observatory (https://vlo.clarin.eu) in order to promote the use of these resources. For this purpose, it is advisable to include descriptive texts to better advertise the data, as users of such catalogues will not necessarily know the background of the project, as they do not access the data via the project website. Such needs may be met in a future release and probably require a change to the CMDI files. At that point, additional metadata files may become necessary to describe the set of resources from a language resources perspective. This is something we currently only do in the case of persistent identifier relations in DOI entries.

Subtasks are parts of the day-long session with one pair of informants where the informants communicate (mostly) without further involvement of the moderator, i.e. a subtask starts when the instructions what to talk about have been completed or the stimulus material has been shown. Some story-retellings are split into several subtasks for each of the images presented. In iLex, subtasks are identified by their tag ids that are the basis of the file names for the corresponding video clips as well as annotation files. Taking this granularity as the basis of a session in the CMDI multimodal recordings sense has a number of implications:

All sessions have the same (minimal) description of the project.
All sessions recorded on the same day share much of the data, i.e. more or less all except references to video and annotation files and the description of the subtask.
The descriptions of the task are repeated across sessions from different days.

Choice of a CMDI Profile for the Public DGS Corpus

While the component structure of CMDI allows the metadata creator to provide metadata in a format exactly matching the requirements of the project, the data is more valuable for users from outside the project if it uses previously defined standardised terminology. This is why CMDI strongly encourages the use of existing and well-defined components (and hopefully vocabularies). This also applies to the choice of profiles (i.e. combinations of components). A best-practice guide (https://github.com/clarin-eric/cmdi-best-practices/releases/download/1.2.0/cmdi_best_practices.pdf) discusses soft criteria when to re-use existing structure vs. defining new. In the case of the minimal metadata to go with release 3 of the Public DGS Corpus it seems reasonable to rely on existing profiles. The lat-session (clarin.eu:cr1:p_1407745712035) is a straight-forward adaption of the IMDI session and comes in a variant for sign language data, lat-SL-session (clarin.eu:cr1:p_1417617523856), building on the results of a workshop on sign language metadata in Nijmegen in 2003 (Crasborn & Hanke, 2003). However, we consider most of the sign language specific parts of this extension inappropriate for metadata on publicly available data, so we follow the recommendations to use a format that avoids lots of fields left to "Unspecified" and use the general lat-session instead. This, however, leaves us with the problem of filling information in a schema not perfectly suitable for sign language data. We discuss the details in the following section. (There is an alternative profile that focuses more on technical aspects, named MultimodalSessionProfile (clarin.eu:cr1:p_1381926654659) which partially poses similar but different problems. Our choice was determined by the fact that many metadata published use lat-session.)
One of the major problems when working with the CMDI approach is the termination of the ISOCat project. The ISOCat provided categories and related vocabularies with multilingual labels and definitions. On the components level, the CMDI Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry) closes this gap, although not with respect to multilinguality. Many limited-size closed vocabularies are moved into the profile xsd documents, but even then the xsd documents lack the definitions previously available via ISOCat. Progress on vocabularies (https://vocabularies.clarin.eu/clavas) seems slow. Moreover, many components still refer to the ISOCat categories although the corresponding web pages are no longer available. For our task, we have tried to access older definitions where available in order to discuss our mapping decisions.

As we expect most, if not all, researchers interested in metadata to be capable of working with English terminology, we do not provide German versions of the CMDI data.

Mapping Options

lat-session::Name and lat-session::Title

The xsd explains these as short vs. full name of the resource. There is no id field for sessions.
Internally, we have

- transcript names: names for the day-long sessions, also reflecting the region, e.g. dgskorpus_ber_01 with ber standing for Berlin,
- subtask names: names for each of the different subtasks, available in in German and English,
- iLex ids for the subtasks.

We use the iLex subtask ids as file names and short titles (lat-session::Name) and provide full names (lat-session::Title) as used in the web interface: «transcript name»: «English name of the subtask». Other projects use a similar strategy, but add "from «Project name»" to the full title.

lat-session::Date
Creation date of the resource: With the resource both containing primary (videos) and secondary data (annotation), this needs some interpretation. As in our environment annotation is a continuous process with no fixed completion date, it only seems natural to consider the recording date as the creation date of the mixed-date resource.

lat-session:Description
For the current purpose of the data, there seems to be no need to fill this field.

lat-session::Location(::Continent/::Country/::Region/::Address/::geoCoordinates)
This can only be interpreted as the location where the recording took place, although the specific address only seems relevant for documentation purposes, without any relevance for interpreting the other data. No explanation could be found why only one address can be provided, but multiple geoCoordinates (points). In the case of our project, the data collection region (the region that both informants are from) always matches the region in which the recording took place. The one exception is the data collection region sh (Schleswig-Holstein) for which the recordings took place in Hamburg (the neighbour region). It would only be a slight mis-use to use this the lat-session::Location slot to describe the data collection region. In that case, ::Continent and ::Country are constantly set to Europe (closed vocabulary value) and Germany (best guess as no vocabulary available), Address, being optional anyway, is left out. Region is filled with the textual description of the data collection region.

lat-session::Project(::Name/::Title/::Id)
Here our problem is what to consider the project. All the recordings to be published form the Public DGS Corpus, so it would be logical to also consider this the project to be listed here. However, the Public Corpus is only one activity of the third-party funded project DGS-Korpus which is only the short title, the full form of which does not even contain the phrase “DGS-Korpus”. While the ::Project::Title might allow both levels to be mentioned in a bulky combined format, this could be considered an inconsistency with ::Project::Name. In the view of longterm usability of the data, it seems best to refer to an entity that hopefully remains a known fact at the institution. We therefore set both ::Name and ::Title to DGS-Korpus, and use “dgs-korpus.de” for the ::Id. In the data in the VLO, we could not identify any majority interpretation how this field should be filled.

lat-session::Content::Genre [multiple]
It seems that there was a vocabulary for this in ISOCat. This is our reconstruction based on http://www.sfs.uni-tuebingen.de/nalida/images/isocat/isocat_hierarchy.html and https://svn.clarin.eu/SMC/trunk/SMC/data/isocat_profile5.rdf.xml:

	interpretingAudience	
	interpretingSource	
	interpretingTarget	
2608	Discourse	The content consists of the spoken/signed utterances of one or more actors. They are produced with the purpose of communicating some thought or intent to the interlocutors present to the event.
2609	Drama	The content is a fictional play that is acted on stage or for broadcasting.
2611	InstrumentalMusic	Instrumental Music
2601	Literature	The content narrates an imaginary event and is valued for its beautiful language.
2605	NewspaperArticle	The content is non-fictional distributed via a newspaper, a magazine or the internet.
2610	PersonalNotes	The content is a brief record of facts or thoughts that act as a mnemonic aid.
2602	Poetry	The content is composed in verse or some similar pattern.
2604	PopularFiction	The content narrates an imaginary event that appeals to popular tastes.
2605	RitualOrReligiousText	The content is concerned with the performance of religious rites consisting of prescribed discourse types.
2600	SecondaryDocument	The content refers to, or comments on, a piece of primary data.
2603	Singing	The content is performed to a tune.
2612	Stimuli	A resource created for a purpose of an experiment

2607	TvRadioFeature	The content is non-fictional spoken/signed text that is broadcast via TV, radio or the internet.

As it seems that this definition of genre is related to written text resources, it is not surprising that one finds this field filled with other values. As all of our data seem to fit under the label "Discourse", we use that throughout all sessions.

lat-session::Content::SubGenre [opt multiple]

Again, no specific advice how to fill this field could be found, and other datasets that we found either do not use this field or they use it in a project-specific way. As our sessions are defined by the task, it does not seem necessary to replicate task-related info into this field, so we leave it out.

lat-session::Content::Task [opt]

In this case, the CLARIN Concept Registry provides some example including "travel-planning" and "frog story" (http://hdl.handle.net/11459/CCR_C-2500_a16d939a-58e3-121d-aaa3-05237ec2d206). From the examples one can conclude that names are to be used here instead of full descriptions. A fuller description might be a candidate text to go into lat-session::Description. So we fill this field with any of the following labels of the 22 subtasks that occur in the Public Corpus:

Experience Report
Deaf Events
Experience of Deaf Individuals
Process Description
Pear Story
Discussion
Fire Alarm
Free Conversation
Frog Story
Regional Specialities
Funny Story
Travel Story
Subject Areas
Warning and Prohibition Signs
Route Description
Calendar Task
Father and Son
Sylvester and Tweety
New vs. Old Signs
Movie Retelling: "Signs"
Joke

lat-session::Content::Modalities [opt]

In this case, the CLARIN Concept Registry contains the following text under example@en (http://hdl.handle.net/11459/CCR_C-2490_44bc38a3-1799-4149-c791-40ac0176f0ff):

"Unknown; Unspecified; speech; writing; gestures; pointing-gestures; signs; eye-gaze; facial-expressions; emotional-state; haptic; song; instrumental music; (source: CLARIN)"

Multiple occurrences are not allowed here, so everything applicable needs to be condensed into one string. In our case, the optimal solution for the time being seems to be

gestures; pointing-gestures; signs; eye-gaze; facial-expressions

lat-session::Content::Subject [opt]

In this case, the CLARIN Concept Registry contains the following text under example@en (http://hdl.handle.net/11459/CCR_C-6543_92dacb75-3098-0d5a-0665-94ddb6274f8c):

"The topic of a news article".

This seems to be at odds with what the notes at http://www.sfs.uni-tuebingen.de/nalida/images/isocat/isocat_hierarchy.html from December 2011 (early phase of ISOCaT) say: At the same position in the element sequence, they list Linguistic Subject instead of Subject and provide as potential values a number of linguistic sub-disciplines. For our purpose, the closest match would be:

2648 text and corpus linguistics

This, however, would be redundant for data from a linguistic corpus. So the decision for the time being is to leave out this field.

lat-session::Content::CommunicationContext

For the fields under lat-session::Content::CommunicationContext it was possible to retrieve most of the definitions from ISOCat from https://svn.clarin.eu/SMC/trunk/SMC/data/isocat_profile5.rdf.xml. Here the definitions are quite useful to understand the granularity in the options suggested:

::Interactivity [opt] Unknown/Unspecified/interactive/non-interactive/semi-interactive

2613	Speech events consists of verbal interaction between at least two Actors.
2614	Speech/song produced without expecting extended verbal responses from hearer(s).
2615	Primarily monologic speech punctuated by repeated interjections from the hearer(s).

::PlanningType [opt multiple] Unknown/Unspecified/spontaneous/semi-spontaneous/planned

2663	Unprompted speech/song
2664	Prompted speech/song
2665	The speaker prepares in detail the structure and content of his/her "performance" in advance

::Involvement [opt multiple] Unknown/Unspecified/elicited/non-elicited/no-observer

2616	Investigator asks speaker(s) to produce isolated phonemes/ words/ utterances / grammatical structures.
2617	The researcher does not interfere verbally with the speech event (other than the researcher's mere presence).
2618	No outside observer is present.

::SocialContext [opt] Unknown/Unspecified/Family/Private/Public/Controlled environment/Public (school)/Community

2619	The access to the communication event is restricted to relatives.
2620	The access to the communication event is restricted to specific individuals of the social environment.
2621	The access to the communication event is allowed to whoever, in a free or in a regulated manner.
2622	The access to the communication event undergoes the agreement to elicit a linguistic behaviour.

::EventStructure [opt multiple] Unknown/Unspecified/Monologue/Dialogue/Multilogue/Not a natural format/Conversation

2659	Communication event with only one main participant.
2660	Communication event between two participants.

2662	Sessions where the number of participants does not define the structure of the communication event.
2661	Communication event with more than two participants.

::Channel [opt] Unknown/Unspecified/Face to Face/Experimental setting/Broadcasting/Telephone/wizard-of-oz/Human-machine dialogue/Other

2593	The transmission of the message ensures full multi-sensorial interaction between speaker and listener(s)
2594	A transmission of the content taking place within a controlled environment for the purpose of testing hypotheses.
	…

This seems to suggest – somewhat unexpectedly – that all sessions, independent of what subtask they reflect, get the following values:

::PlanningType = spontaneous
::Involvement = non-elicited
::SocialContext = Controlled environment
::EventStructure = Dialogue
::Channel = Face to Face

Only for ::Interactivity it remains unclear whether all our formats fall under either interactive or semi-interactive or if the value non-interactive is to be used as well. The definition of "non-interactive" refers to the expectation of extended feedback. As it is difficult to know from just observation what the conversation participants had in mind, this can only refer to the task designers' expectations. This means, however, that the decision cannot be taken by observing any means of measuring interactivity in the actual conversation, but only how conversations were planned. Actually, we have documented our expectations on the degree of interactivity in a binary fashion. Basically, this is "semi-interactive" for jokes and retellings (*) and "interactive" (**) in all other cases:

Experience Report **
Deaf Events **
Experience of Deaf Individuals **
Process Description *
Pear Story *
Discussion **
Free Conversation **
Fire Alarm *
Frog Story *
Regional Specialities **
Funny Story **
Travel Story *
Subject Areas **
Warning and Prohibition Signs **
Route Description **
Calendar Task **
Father and Son *
Sylvester and Tweety *
New vs. Old Signs **
Movie Retelling: "Signs" **
Joke *

lat-session::Content::ContentLanguage [opt mult]

It seems obvious that DGS should be listed here. This becomes less clear after a look at the fields to accompany the language: ::Dominant, ::SourceLanguage, ::TargetLanguage. Here one would expect the choice between true and false for ::Dominant, in our case always to be answered with "true", and ::SourceLanguage and ::TargetLanguage to be left out except in the context of interpreting. However, http://hdl.handle.net/11459/CCR_C-2468_e4135e12-c272-171e-a8a2-48339228387b lists "en_UK; fr_FR; de_SW (source: CLARIN)" as an example, and all three fields are non-optional. We assume that this an error in the session definition when transferred from IMDI format. The best solution seems to ignore the example and fill ::Dominant with true (which would be redundant unless a second language is listed) and to specify "Unspecified" for source and target language.

A more interesting question is when to list other languages:

First, there are some cases in the corpus that can clearly labelled as LBG (Sign-Supported German) instead of proper DGS. However, there is a continuum between DGS and LBG. At what point would one want to flag this? And does that global label not stigmatise the informants of that session even if it might only apply to one of them?

Second, some discussions cover other sign languages, sometimes quoting signs from those languages. While this is covered in the annotation, should this also be indicated in the metadata, even if it is only one out of several thousand tokens?

The decision taken here is a pragmatic one: As we want metadata to remain independent of annotation as annotation changes over time, we only list DGS in the metadata.

lat-session::Actor::Role

Here it is the IMDI xsd providing an open vocabulary, namely "consultant, contributor, interviewer, researcher, publisher, collector, translator" (https://www.mpi.nl/IMDI/Schema/IMDI_3.0.xsd) while several CMDI data sets use the value "speaker/signer". Here we mix from sources and use "speaker/signer" for the two informants and "interviewer" for the moderator. Researchers, publishers etc. are not listed.

lat-session::Actor::Name and ::FullName [opt] and ::Code

The CLARIN Concept Registry gives a definition: "The name of the person participating in the content of the recording as it is used by others in the transcription. (source: CLARIN)" (http://hdl.handle.net/11459/CCR_C-

2557_362261de-9ef9-1fbc-a5cd-5d55a8d44aa2). This definition does not apply in the context of sign language corpora, but is an obligatory field. The full name is well-defined, but we are not prepared to make names public, so this optional field is left out. For the time being, we set both ::Name and ::Code to the informant codes used in the project and in some annotation files, namely a 2-or-3-letter abbreviation of the data collection region followed by a two-digits running number, not to be confused with the transcript code which look almost identical in format. There is no way to infer from transcript code to informant code or vice versa.

lat-session::Actor::FamilySocialRole and ::EthnicGroup and ::BirthDate and ::Education

All these fields are mandatory in the profile, but we definitely do not want to report such data, so we have to fill them with "Unspecified".

lat-session::Actor::Sex

As none of the participants in the corpus considered themselves non-binary at the time of filling out the questionnaires, filling out this field is straight-forward.

lat-session::Actor::Age

The age can be specified as an estimated age, an exact age, or an age range. As said in the introduction, we only make age ranges publicly available. As our age ranges are defined in years, we only provide the years field and leave out the optional months and days data. For some reasons, an upper limit, called MaximumAge is mandatory for a range. So our fourth age group 61+ needs to be coded as 61-999.

lat-session::Actor::ActorLanguage

In addition to the language name and ISO639-3 code, this specification of language has mandatory elements ::MotherTongue and ::PrimaryLanguage both allowing the values "Unspecified", "Unknown", "true" and "false". We leave both cases set to "Unspecified" for data privacy reasons, so do not need to decide how to fill the concept of ::MotherTongue for persons with sign language as their first language at this point of time.

lat-session::Resources::MediaFile [opt mult]

In our case, it is obvious how to fill ::Type and ::Format (with "video" and "video/mp4" respectively). The other fields, however, are not that obvious:

lat-session::Resources::MediaFile::Size and ::TimePosition

http://hdl.handle.net/11459/CCR_C-2580_6dfe4e09-1c61-9b24-98ad-16bb867860fe offers a definition for size: "The size of the resource with regard to the SizeUnit measurement in form of a number. (source: CLARIN)". However, one often finds values such as "18.5 MB" which seems not to be compatible with that definition. The reason probably is that lat-session does not contain a SizeUnit element (nor does it allow multiple Size elements). So data without a size unit, as is also used in one of the CMDI examples (https://raw.githubusercontent.com/wiki/clarin-eric/cmdi-toolkit/examples/example-phonological-corpus-1_2.cmdi), seems ambiguous:

```
<cmdp:Format>audio/x-aiff</cmdp:Format>
<cmdp:Size>1134212</cmdp:Size>
```

Typically, researchers are not very interested byte size, caring more about amounts of information contained (e.g. number of words, duration of dialogue, etc.). Nevertheless, the extra elements TimePosition::Start und ::End allow us to specify the size of the resources as a duration, although they might have been thought to refer to a clipping from a larger resource. So we fill ::Size as the size of the file in bytes, TimePosition::Start with "00:00:00:00" and TimePosition::End with the duration of the video.

lat-session::Resources::MediaFile::Quality [opt]

This field allows to define quality in terms of a range 1–5. Apparently, this only makes sense to provide relative quality measures within one corpus. Even then it remains problematic to decide what aspects this quality measure has in mind: Ease of viewing for the human reader, suitability for image processing, or any combination? For the time being, we do not fill out this field.

lat-session::Resources::MediaFile::RecordingConditions [opt]

The only other field available on MediaFiles is the optional ::RecordingConditions. http://hdl.handle.net/11459/CCR_C-2566_5a4ee887-bc58-38ee-9b1e-a06f1916d63c gives a definition ("Description of the technical conditions under which the resource was recorded. (source: CLARIN)" as well as an example: "Microphone Sennheiser MD60 (source: CLARIN); sound-proof cabin (source: CLARIN); capture

device USB MAUDIO 4300 (source: CLARIN)". More important than the brand and type of camera used are the recording resolution and interlaced vs. progressive as well as the camera perspective.

lat-session::Resources::MediaFile::Access [opt]

Finally there is another complex element under the node ::MediaFile, namely ::Access. It gives information how the actual data (as opposed to the metadata) can be accessed. In the case of the DGS Public Corpus, this seems superfluous: The actual data are available from the same website as the metadata. This will become an issue, however, if metadata is harvested to be included in catalogues. For our case, a persistent identifier to the data would suffice, but licensing conditions should be documented here or elsewhere.

lat-session::Resources::WrittenResource [opt mult]

In the case of our sign language corpus, WrittenResource refers to annotation files. We provide a record for each format made available.

lat-session::Resources::WrittenResource::Date [opt]

As we understand annotation as an ongoing process, we do not provide any date.

lat-session::Resources::WrittenResource::Type

http://hdl.handle.net/11459/CCR_C-3900_8cbd76e3-556e-8271-1baf-b2170c7017ab provides the following text:
"example@en: primary text, annotation, etnography, study, etc.
scopeNote@en: This data category allows specifying the type of written resource such as text, annotation, lexical research, transcription, etc."
In our terminology, annotation is the wider concept when compared to transcription. As our annotation files also include translations, annotation is the term to be used.

lat-session::Resources::WrittenResource::SubType [opt]

http://hdl.handle.net/11459/CCR_C-3901_af291d73-7e01-a844-b88e-0d3f0c95bd84 provides the following text:
example@en: dictionary, terminology, wordlist, lexicon, etc. (if written resource type, i.e. the superordinate type, is LexicalAnalysis). (source: IMDI)
scopeNote@en:This element allows to specify the sub-type of a written resource. Different types of written resources have different controlled vocabularies for sub-types: the type 'Lexical research', for instance, has as sub-type-vocabulary {dictionary, terminology, wordlist, lexicon, … }. In case the written resource type is annotation, the sub-type specifies the type of annotation, such as phonetic, morphosyntax, etc. (source: IMDI)"
As the profile does not allow multiple attributions, it seems best to leave this field out.

lat-session::Resources::WrittenResource::Format

While http://hdl.handle.net/11459/CCR_C-2465_4444eb51-7cf7-0ff7-7687-7f741f3a4f84 suggests contents such as "Zip", but formats such as "text/x-eaf+xml" as found in sample data are much more convincing.

lat-session::Resources::WrittenResource::Size

Analogous to the case ::MediaFile::Size.

lat-session::Resources::WrittenResource::Derivation [opt]

http://hdl.handle.net/11459/CCR_C-2518_ea48054d-f23f-c493-bcc0-067561b87c67 lists Unknown, Unspecified, Original, Analysis, Translation, Commentary, Criticism, Annotation as examples. Unfortunately, multiple values such as Translation + Annotation are not possible. Even if this seems redundant with ::Type, the value "Annotation" seems to be the best fit.

lat-session::Resources::WrittenResource::LanguageId [opt]

For the html and srt version of the annotation data, we provide separate versions in German and English. In these cases, the written resources are tagged with the language. In the other cases, the annotation file is multilingual. There we do not provide the LanguageId.

lat-session::Resources::WrittenResource::Anonymized [opt]

We set this field to true in all cases.

lat-session::Resources::WrittenResource::Validation [opt]

So far, none of our quality control measures are formalized to a degree that they would be describable here. As a consequence, this field is not included.

References

Crasborn, Onno / Hanke, Thomas (2003). Metadata for sign language corpora. Manuscript. http://sign-lang.ruhosting.nl/echo/docs/ECHO_Metadata_SL.pdf