HZSK core metadata set

The HZSK core metadata set is a collection of metadata that is partly manually added by the user and partly automatically created by Coma. Coma provides the structure of the metadata (containers/data types rendered in italics, elements in lower-case/camelCase, attributes with preceding "@") and some automatically created mainly technical information on the corpus resources (rendered in grey). These HZSK conventions apply to values of Coma attributes and elements entered by the user, and attributes and values in the Coma descriptions.

The HZSK Catalogue metadata set is a subset of the HZSK core metadata set that consists of all DC and OLAC metadata items and the obligatory items of the core set. For each corpus, this smaller metadata set is the basis for the CMDI files that are made publicly available for metadata harvesting via OAI-PMH. The "obligatory" automatically generated items (rendered in grey) are not considered. The catalogue metadata should make sure the resources are discoverable for the research community and help potential corpus users decide whether to request access or not. Obviously, this smaller set is also restricted to metadata that allow the participants to remain completely anonymous. Our CMDI profiles and components are however more flexible when it comes to obligatory elements than these conventions, since legacy corpora might always include exceptions, e.g. subjects for which the mother tongue is unknown or communications where even the year of recording has for some reason been lost. This information can unfortunately not always be reconstructed, but such corpora should still validate against our CMDI scheme. We also provide some additional (mainly technical) information in our CMDI profiles and express information according to commonly used components, sometimes leading to redundancy. Since we always generate complete CMDI instances from Coma files, this is not a consistency problem.

General conventions for attributes and values

We use English for all attribute names of the HZSK conventions. We capitalize the first word only of the attribute and write all values in lower case, apart from words which always require upper case, such as geographical or language names etc. If a corpus was originally described in another language, we use the values in this language together with our English attribute names, that is we use the translate original attribute names for which there are corresponding HZSK conventions. We do not however attempt to translate the values, e.g. information on a speaker's education or occupation, into English. We neither translate information that is obviously theory laden - the attribute "room" simply referring to the room in which a communication took place could be translated, but not the "institutional place".

Though this can not be described as part of the conventions below, we also make sure all corpus design specific attributes are explicitly encoded. Sometimes, the parameters of the corpus design are only recognizable as names of files or in speaker sigles, and in this case they are often only decipherable to original project members. Some of this information will be encoded anyway as part of the conventions, e.g. the age of speakers in a corpus with different age groups, or the first and second language of speakers. This leads to a certain redundancy, but we still prefer to keep or create corpus design specific attributes to facilitate the use of the corpus as it was intended.

Multiple and complex values

Some metadata elements can have multiple values, e.g. the Language element of the Dublin Core. According to the DC conventions, these elements should be repeated. Since the use of the same Key in a description repeatedly is not possible in Coma, all values are listed together in one value field, separated by semicolon. The DC item “Language” is encoded “DEU; POR; SPA; TUR” for the corpus DiK in German, Portuguese, Spanish and Turkish. For complex values, commas are used for the first level, and slashes for the second. The DC item “Publisher” thus has a value such as “Hamburger Zentrum für Sprachkorpora, Max-Brauer-Allee 60 / D-22765 Hamburg, corpora@uni-hamburg.de” with the first part being the person, the second the address with its parts separated by a slash, and the third the email address.

Controlled vocabularies and keywords

When possible, we included ISO standard vocabularies for codes and names of languages or countries. For some elements, a closed controlled vocabulary (CCV) exists. This means that only the listed values are accepted. Sometimes it is possible to specify a standardized domain for the values of a metadata element, for example language and country codes that are ISO-standardized. If we were to extend this domain with values such as “unknown” or “unspecified”, it would no longer be easily describable and referable. In these cases we rather omit the entire metadata element. For some elements, an open controlled vocabulary (OCV) exists. In these cases the existing values should be used when possible, and otherwise the new value has to be added to the vocabulary. The least restricting option is called “keyword control” and means that the value should use specific words, e.g. if “television” is listed as keyword, the value of the “type” attribute of a Communication should not be “TV broadcast”, since this makes searching and filtering much more difficult.

Overview of the metadata set

Below, all relevant Coma metadata items are listed as an overview. However, this is not a documentation of the Coma XML schema, as e.g. the quantifiers of the HZSK conventions differ from the default ones. Below this general overview, the items for which there are specific non-self-explanatory HZSK conventions are described in more detail. For documentation on the usage of the remaining items, we refer to the general Coma documentation. For documentation on the usage of the OLAC and DC items, we refer to the respective documentation.

corpus
@id (GUID)
@name (string)
[@type] (string)
@uniqueSpeakerDistinction (XPath)
@schemaVersion (string)
  description
    DC:contributor (string)
    DC:coverage (string)
    DC:created (date)
    DC:creator (string)
    [DC:date] (date)
    DC:description (string)

DC:format (string)
DC:identifier (string)
DC:language (ISO 639-3)
DC:publisher (string)
[DC:relation] (string)
DC:rights (string)
DC:rightsHolder (string)
[DC:source] (string)
DC:subject (string)
DC:title (string)
DC:type (string)
HZSK:keywords (OCV)
HZSK:shortdescription (string)
OLAC:compiler (string)
OLAC:data-inputter (string)
OLAC:developer (string)
OLAC:researcher (string)
OLAC:sponsor (string)
*associatedFile**
@id (GUID)
@name (string)
@type (OCV)
file+
@id (GUID)
@name (<filename>.<extension>)
[@type] (string)
[description]
mimetype (string)
url (anyURI)
description
Language (string, keyword control)

communication
@id (GUID)
@name (string)
@type (string, keyword control)
description
[Background information] (string)
Project name (string)
[Source] (string)
placeInTime
@type="Communication"
[@name] (string)
location
[city] (string)

country (ISO-3166-1 English country name)
latitude (string)
longitude (string)
locationPrecision (CCV)
period
periodStart (dateTime)
[periodDuration] (long)
periodPrecision (CCV)
[description]
[Precision] (string)
[Region] (string)
language+
@type="Communication"
@name (ISO 639-3 English language name)
LanguageCode (ISO 639-3)
[description]
*recording**
@id (GUID)
@name (string)
[@type] (string)
recordingDuration (long)
recordingDateTime (dateTime)
[description]
[Recording person] (string)
[Recording device] (string)
file+
@id (GUID)
@name (<filename>.<extension>)
@type (CCV)
[*description*]
mimetype (string)
url (anyURI)
*transcription**
@id (GUID)
@name (string)
[@type] (string)
file+
@id (GUID)
@name (<filename>.<extension>)
@type="EXMARaLDA segmented transcription"
description
[Segmentation algorithm] (CCV)
<Segmentation algorithm>:<Unit>+ (int)
e (int)
sc (int)

EXB-SOURCE (anyURI)
mimetype (string)
url (anyURI)
*file**
@id (GUID)
@name (<filename>.<extension>)
@type (CCV)
[description]
mimetype (string)
url (anyURI)
description
Alignment status (CCV)
Transcription status (CCV)
Transcriber[, <language>] (string)
[Transcription checker[, <language>]] (string)
Transcription convention (string)
[Transcription date] (string)
[Transcription number] (string)
[Transcription quality] (string)
[Annotation status] (CCV)
[Annotation type: <tier category>] (string)
[Annotator[, <tier category>]] (string)
[Annotation checker[, <tier category>]] (string)
[Translation status] (CCV)
[Translator[, <language>]] (string)
[Translation checker[, <language>]] (string)
*annotation**
file
@id (GUID)
@name (<filename>.<extension>)
[@type] (string)
description
Annotation type: <tier category> (string)
mimetype (string)
url (anyURI)
*associatedFile**
@id (GUID)
@name (string)
@type (OCV)
file+
@id (GUID)
@name (<filename>.<extension>)
[@type] (string)
[description]
mimetype (string)

url (anyURI)
description
Language (string, keyword control)

speaker
@id (GUID)
[@name] ([<firstName>] [<lastName>])
@type (CCV)
sigle (string)
sex (CCV)
[description]
[Background information] (string)
language+
[@type] (string)
@name (ISO 639-3 English language name)
LanguageCode (ISO 639-3)
[description]
placeInTime
@type="Birth"
[@name] (string)
location
[city] (string)
country (ISO-3166-1 English country name)
latitude (string)
longitude (string)
locationPrecision (CCV)
period
periodStart (dateTime)
periodDuration="0"
periodPrecision (CCV)
[description]
[Precision] (string)
[Region] (string)
*placeInTime**
@type="Education"
[@name] (string)
[location]
[city] (string)
country (ISO-3166-1 English country name)
latitude (string)
longitude (string)
locationPrecision (CCV)
[period]
[periodStart] (dateTime)
[periodDuration] (long)

[periodPrecision] (CCV)
[description]
[Precision] (string)
[Region] (string)
[Degree] (string)
*placeInTime**
@type="Occupation"
[@name] (string)
[location]
[city] (string)
country (ISO-3166-1 English country name)
latitude (string)
longitude (string)
locationPrecision (CCV)
[period]
[periodStart] (dateTime)
[periodDuration] (long)
[periodPrecision] (CCV)
[description]
[Precision] (string)
[Region] (string)
*associatedFile**
@id (GUID)
@name (string)
@type (OCV)
file+
@id (GUID)
@name (<filename>.<extension>)
[@type] (string)
[description]
mimetype (string)
url (anyURI)
description
Language (string, keyword control)
role
@type=”#participant”
[description]
[language]

Documentation and guidelines

corpus

The attribute *name* contains the full name of the corpus with the common abbreviation in parentheses.

For the *HZSK:keywords*, the keywords listed at http://www.corpora.uni-hamburg.de/sfb538/en_overview.html are used. If necessary, further keywords may be added, but there should be no synonyms to existing keywords.

All *HZSK:shortdescription* values should include the following information (where applicable):

- audio/video recordings
- speech type(s)
- bilingualism of the speakers: languages and type
- age of the speakers
- communication types
- languages used in communications
- recording design: how many, how often and for how long
- corpus design specific attributes
- available material (tasks, tests etc.)

communication

The attribute *type* of the communication is used to encode the communication type using existing keywords.

Background information as a key in the description is used for all information on the situation and the time before the communication.

Source as a key in the description is used for communications that were not recorded but sampled from radio broadcasts etc.

transcription

The transcription name should be the same as the recording's to which it belongs, if the relationship is 1:1.

The description key *Alignment status* has the following CCV: ("not aligned"|"partly aligned"|"fully aligned"|"unknown")

The description key *Transcription status* has the following CCV: ("not transcribed"|"partly transcribed "|"fully transcribed "|"unknown")

The description key *Transcription convention* might differ from the segmentation algorithm used. For example, many transcription conventions based on standard orthography and punctuation can be segmented with the HIAT segmentation algorithm.

The description key *Transcription number* is only relevant when one communication has been split up in several transcriptions.

The description keys for annotation and translation status are only relevant when the transcriptions of the corpus have been systematically annotated and translated. There is no need to describe all transcriptions as "not annotated" and "not translated" by default. The CCVs for

these keys are analogue to those for alignment and transcription status:
Annotation status ("not annotated"|"partly annotated"|"fully annotated"|"unknown")
Translation status ("not translated"|"partly translated"|"fully translated"|"unknown")

The description key *Annotation type: <tier category>* is repeated for all used tier categories for A(nnotation)-type tiers in the EXMARaLDA basic transcription with the description of the annotation type as value.

The values of the attribute type are restricted by the allowed transcriptionFileTypes in Coma: ("EXMARaLDA basic transcription"|"EXMARaLDA segmented transcription"|"TEI transcription"|"FOLKER transcription"). For each transcription, at least a file of the type "EXMARaLDA segmented transcription" is required.

The description key *Segmentation algorithm* only applies to segmented transcriptions and has the following CCV:
("GENERIC"|"HIAT"|"DIDA"|"GAT"|"cGAT_MINIMAL"|"CHAT"|"CHAT_MINIMAL"|"IPA")

annotation

The annotation name should be the same as the transcription's to which it belongs, if the relationship is 1:1.

The description key *Annotation type: <tier category>* is repeated for all used annotation types of the SEXTANT annotation file with the description of the annotation type as value. The annotations from the SEXTANT file can be integrated into the segmented transcription, but the description of the annotation types should remain here to indicate the origin of the annotations.

recording

The recording name should be the same as the transcription's to which it belongs, if the relationship is 1:1.

One recording contains all media files derived from one and the same recording, even if these are both audio and video files.

The *type* attribute of the file element in the recording is used to differentiate between audio and video files, ("audio"|"video").

associatedFile

Associated files can be attached to the corpus, to communications or to speakers. Materials that apply to all communications, e.g. a task description, are described as corpus materials, whereas communication or speaker specific materials are attached to communications and speakers.

The *type* attribute is used with existing values/keywords when possible.

For *Language* as a key in the description, the ISO 639-3 English language names are used, with multiple languages separated by semicolon.

placeInTime

The *placeInTime* element provides elements that indicate the precision of location and period information:
locationPrecision ("address"|"street"|"city"|"state"|"country")
periodPrecision ("year"|"month"|"day"|"hour"|"millisecond")
The information that is not exact still has to be filled in for the *period* element, in this case e.g. month and day are simply set to “01”. Additionally, a description key *Precision* can be used if more information is needed, e.g. one would like to add if the age of a speaker was given as “40-60” years or “mid 50s” and then encoded somehow in accordance with this information.

speaker

For the *type* attribute of the speaker, there is a CCV ("subject"|"researcher"|"other"|"unknown") making sure contributions from subjects can be evaluated separately.

The element *sex* is restricted by Coma ("male"|"female"|"unknown").

The *type* attribute of speaker languages should be used to indicate whether this is a speaker’s L1 etc. - if this is possible to decide.

All information on the speakers present and former education and occupation is encoded as placeInTime elements with one element for each entry and the attribute *type* set to either Education or Occupation. There is often information on the highest (academic) degree of speakers - this information is encoded within the Degree key in the description of an Education placeInTime element.