Metadata Guide for Linguistics Data

Metadata documentation

For more information, please contact:

swissubase@ub.uzh.ch

FORS explore.understand.share.

UNIL | Université de Lausanne

Universität Zürich[UZH]

Version:	1.2	Persistent link:	n/a
Last modified:	21.11.2024	*License:*	CC BY
Introduced:	04.07.2022	*Publisher:*	Language Repository Switzerland (LaRS) swissubase@ub.uzh.ch

Preamble: This guide provides a description of the linguistics-specific metadata fields which describe the dataset and data files.

> Please note the input rules to enter metadata in the third column.

> Examples can be found in the description column and under examples.

Table of contents

1. RESOURCE METADATA

Field name EN	Description EN	Input rule	Examples
Resource type	The Resource type describes the main focus of the dataset.	n/a	Corpus
Resource description	Description of the resource/dataset	n/a	n/a
Keywords	Keywords to classify the resource	Comma separated	Keyword 1, keyword group 2, keyword 3
Validation information	Enter the dataset's validation info		
Participants	Use this section to provide general details about the participants in your project, e.g. number of participants, gender, age range, etc	n/a	n/a
Participants.Number of persons	Number participants in the research	n/a	12
Participants.Gender	Gender of the participants	n/a	Female
Participants.Age range (Start)	Start of the age range of the participants	n/a	24
Participants.Age range (End)	End of the age range of the participants	n/a	40
Participants.Birthdates	Birthdates of the participants	Comma separated; - Single: XXXX - Range: XXXX-XXXX - Before: before XXXX - After: after XXXX	1980-2000 before 1990 1984, 1990, 1992

		- Multiple single dates: XXXX, XXXX	

1.1 Controlled Vocabulary

Field name EN	Controlled Vocabulary Values	Description
Resource type	Corpus	For data that can be considered text, speech, video, multimodal/multimedia corpus
Resource type	Experimental resource	Neurolinguistic, Sociolinguistic etc.
Resource type	Computational resource	e.g. Language models
Resource type	Lexical/Conceptual resource	Lexica, ontologies, dictionaries, word lists
Resource type	Language description	e.g. Grammers
Resource type	Other	n/a
Participants.Number of persons	n/a	n/a
Gender	Male	n/a
Gender	Female	n/a
Gender	Mixed	n/a
Gender	Unknown	n/a

Gender	Other	n/a
Age range (Start)	n/a	n/a
Age range (End)	n/a	n/a

2. LANGUAGE METADATA (TO ADD ONE BLOCK IS MANDATORY)

Field name EN	Description EN	Input rule	Examples
Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/media/audio/interviews/
Language name	The name and ISO-Code (639-3) of the language.	n/a	por - Portuguese
Linguality type	The linguality type defines whether a data collection/resource/speaker is monolingual, bilingual or multilingual. If a resource is bilingual or multilingual, you can add an additional language metadata block for the other language(s).	n/a	monolingual
Country	The country to which the language resource refers (usually a country where the data were collected).	n/a	Portugal

Language variety name	The language variety type defines if the language of the described resource is a dialect, jargon or other. E.g. If Portuguese is chosen as the language name, a possible variety name is e.g. Miguelense (for the Portuguese language variety that is spoken on S. Miguel/Açores.)	n/a	Miguelense
Language variety type	The language variety type defines if the language of the described resource is a dialect, jargon or other.	n/a	dialect
Language status	Language status describes the actual current status of a language. E.g. Portuguese is a living language.	n/a	Living
Modality type	Modality type refers to the manifestation of the language: e.g. sign language or spoken language	n/a	spoken language
Naturality	Modality type refers to the manifestation of the language: e.g. sign language or spoken language	n/a	natural

2.1 Controlled Vocabulary

Field name EN	Controlled Vocabulary Values	Description
Language Name	n/a	Format of the list is: ISO Code - Language Name
Linguality Type	monolingual	Resource or speaker of one language
Linguality Type	bilingual	Resource or speaker of two languages
Linguality Type	multilingual	Resource or speaker of three or more languages
Linguality Type	not applicable	n/a
Linguality Type	other	n/a
Country	n/a	n/a
Language variety type	dialect	Regional variety of a language expressed by different pronunciation, grammar and/or vocabulary.
Language variety type	jargon	Social variety of a language expressed by a specific social group that is defined by profession, standing and/or milieu.
Language variety type	other	n/a
Language variety type	not applicable	n/a

Language status	Ancient	Examples for ancient languages are Etruscan, Latin, Gothic or Hittite. The term covers languages that are not any longer in use but are documented.
Language status	Constructed	An example for a constructed language is Esperanto. The term covers languages that are created and constructed on purpose.
Language status	Extinct	Examples of extinct languages are Samaritan Aramaic or Tupí. These languages are not any longer spoken.
Language status	Genetic	An example for a genetic language unit is South Slavic languages. Most of the time the genetic languages term covers a group of languages (here Serbian Bosnian etc. are covered by the term).
Language status	Genetic, Ancient	An example for a genetic ancient language is Prakrit. The term covers a group of ancient languages (here a group of vernacular Middle Indo-Aryan languages).
Language status	Genetic-like	Examples for genetic-like languages are Portuguese-based creoles or pidgins.
Language status	Geographic	The term covers a group of languages which are bound by their common geographical location. An example for geographic languages is Caucasian languages.
Language status	Historical	An example for a historical language is Anglo-Norman.
Language status	Living	A language that is still in use by a community.
Language status	Special	Remark: Not in use according to ISO database (field will be deleted)
Language status	not applicable	n/a

Langauge status	other	n/a
Modality type	body gesture	The term covers non-verbal communication expressed by a body part.
Modality type	facial expression	The term covers non-verbal communication by facial expressions.
Modality type	voice	The term covers any involvement of the voice. If the focus is on spoken language rather than on any other form of voice, please choose spoken language.
Modality type	combination of modalities	The term covers a combination of modalities named in this list.
Modality type	sign language	The term covers communication expressed by sign languages.
Modality type	spoken language	The term covers communication expressed in spoken form.
Modality type	written language	The term covers communication expressed in written form.
Modality type	other	n/a
Modality type	not applicable	n/a
Naturality	assisted	n/a
Naturality	elicited	The term covers a situation in which an utterance was stimulated, evoked or educed (emotions, feelings, responses).
Naturality	natural	The term covers a situation in which an utterance was made in a natural way.

Naturality	planned	The term covers a situation in which an utterance was planned and supposed to be uttered.
Naturality	prompted	The term covers a situation in which an utterance was led towards a specific answer.
Naturality	read speech	The term covers a situation in which an utterance was made by reading a text.
Naturality	semi planned	The term covers a situation in which an utterance was semi planned and semi spontaneous.
Naturality	spontaneous	The term covers a situation in which an utterance was made spontaneously.
Naturality	other	n/a
Naturality	not applicable	n/a

3. ANNOTATION METADATA

Field name EN	Description	Input rules	Examples
Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/media/audio/annotations
Annotation type	Specifies the annotation type of the annotated version(s) of a resource, or the annotation type a tool/service requires or produces as an output	n/a	lemmatization
Annotation tiers	Please insert the annotation tiers.	n/a	interval
Annotation format	The file format in which data annotation was implemented. E.g. application/xml+tei	- Please use media types if possible, in accordance to this list: https://www.iana.org/assignments/media-types/media-types.xhtml - if no media typ is available, please use the full name of the format and specify the commonly used abbreviation in brackets: Format Name (Abbr)	application/xml+tei
Controlled Vocabulary	The controlled vocabulary or tag set that was used for annotation. E.g. Stuttgart-Tübingen-Tagset (STTS)	Please use the full name of the controlled vocabulary or tag set and specify the commonly used abbreviation in brackets: Name of	Stuttgart-Tübingen-Tagset (STTS)

		the Tag set (Abbr), Name of the controlled vocabulary (Abbr)	

4. AUDIO METADATA

Field name EN	Description	Input rules	Examples
Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/media/audio/music
Audio genre	The genre contains classifications of the resource content: E.g. Traditional folk song, political speech	Short meaningful description, max. 10 words	Traditional folk song
Codec	Indication of the encoding method how the audio signals were converted for storage. E.g. flac	n/a	flac
Media type	Standardised naming of the file format, a media type, e.g. audio/mpeg	n/a	audio/mpeg
Duration	The duration of the resource in HH:MM:SS	HH:MM:SS	00:53:04
Duration of effective speech	The duration of the effective speech, the duration without interruptions, pauses etc. in HH:MM:SS	HH:MM:SS	00:40:10

4.1 Controlled Vocabulary

Field name EN	Controlled Vocabulary Values	Description
Codec	FLAC	Free Lossless Audio Codec, supported containers: MP4, Ogg, FLAC
Codec	SHN	Shorten (SHN), a lossless audio compression algorithm, is deprecated.
Codec	MP3	MPEG-1 Audio Layer III, supported containers: MP4, ADTS, MPEG, 3GP; A MPEG-1 Audio Layer III stored in a MPEG container without video tracks is usually called a MP3.
Codec	Vorbis	Vorbis, supported containers: WebM, Ogg
Codec	ATRAC	Adaptive Transform Acoustic Coding, an audio compression algorithm by Sony, is deprecated.

Codec	AAC	Advanced Audio Coding, supported containers: MP4, ADTS, 3GP
Codec	MPEG	Supported Containers are MPEG-1 and MPEG-2
Codec	RealAudio	Codec family for RealAudio files
Codec	ALAC	Apple Lossless Audio Codec, supported containers: MP4, QuickTime (MOV)
Codec	AMR	Adaptive Multi-Rate, supported container: 3GP
Codec	Opus	Opus, supported containers: WebM, MP4, Ogg
Codec	PCM	Raw Pulse Code Modulation Audio as Wave, SND, AU or OGG
Codec	other	n/a
Media type	audio/mid	The usual file extensions are .mid or .rmi
Media type	audio/mp4	The usual file extension is .mp4
Media type	audio/mpeg	The usual file extension is .mp3

Media type	audio/vnd.wave; audio/x-wav; audio/wav; audio/wave	The usual file extension for this is .wav
Media type	audio/x-aiff	The usual file extensions are .aif, .aifc or .aiff
Media type	audio/vnd.rn-realaudio	The usual file extensions are .ra or .ram
Media type	audio/ogg	The usual file extension is.ogg. Ogg can contain different audio streams/codecs in Vorbis, Opus, FLAC or PCM.
Media type	audio/vorbis	The usual file extension is .ogg. Vorbis is often used in combination with ogg-container.
Media type	audio/basic	The usual file extensions are .snd or .au
Media type	audio/x-flac	The usual file extension is .flac
Media type	other	n/a

5. TEXT METADATA

Field name EN	Description	Input rules	Examples
Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/media/transcripts
Media type	Standardised naming of the file format, a media type, e.g. text/xml	n/a	text/xml
Character encoding	Character set used to encode the text resource, e.g. UTF-8	n/a	UTF-8
Provenience/derivation	The origin of the text. Is it a derivation, e.g. a transcript, or is it e.g. original written speech?	Short meaningful description, max. 10 words	Transcribed folk song
Text genre	The genre contains classifications of the resource content, e.g. Folk song about local customs	Short meaningful description, max. 10 words	Folk song about local customs

5.1 Controlled Vocabulary

Field name EN	Controlled Vocabulary Values	Description

Media type	text/csv	Media type for comma separated values, file extension is normally .csv
Media type	text/html	Media type for HTML based documents, the usual file extension is .html
Media type	text/plain	Media type for plain text documents, the usual file extension is .txt
Media type	text/sgml	Media type for Standardized Generalized Markup Language, the usual file extension is .sgml
Media type	text/tab-separated-values	Media type for tab separated values, the usual file extension is .tsv
Media type	text/turtle	Media type for Terse RDF Triple Language files, the usual file extension is .ttl
Media type	application/vnd.xmi+xml	Media type for XML Metadata Interchange files, the usual file extension is .xmi
Media type	application/json	Media type for JSON files, the usual file extension is .json.
Media type	text/xml; application/xml	Media type for XML documents, the usual file extension is .xml.
Media type	application/x.tmx+xml	Media type for Translation Memory eXchange, the usual file extension is .tmx
Media type	application/x-xces+xml	Media type for X Corpus Encoding Standard

Media type	application/tei+xml	Media type for TEI files. The usual file extensions are .tei or .xml
Media type	application/rdf+xml	Media type for RDF files in XML. The usual file extension is .rdf
Media type	application/xhtml+xml	Media type for XHTML files. The usual file extensions are .xhtml, .xht, .xml, .html and .htm
Media type	application/emma+xml	Media type for Extensible MultiModal Annotation markup language in XML
Media type	application/pls+xml	Media type for Pronunciation Lexicon Specification in XML
Media type	application/voicexml+xml	Media type for Voice Extensible Markup Language in XML
Media type	application/x-tex	Media type for TeX files
Media type	text/rtf; application/rtf	Media type for Rich Text Format files,the usual file extension is .rtf
Media type	application/x-latex	Media type for Latex source files
Media type	application/pdf	Media type for PDF files, the usual file extension is .pdf
Media type	application/x-msaccess	Media type for MS Access databases, the usual file extensions are .mdb or .accdb
Media type	application/msword	Media type for old Microsoft Word files, the usual file extension is .doc

Media type	application/vnd.openxmlformats-officedocument.wordprocessingml.document	Media type for Microsoft Word files, the usual file extension is .docx
Media type	application/vnd.ms-excel	Media type for old Microsoft Excel files, the usual file extension is .xls
Media type	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	Media type for Microsoft Excel files, the usual file extension is .xlsx
Media type	other	n/a

6. VIDEO METADATA

Field name EN	Description	Input rules	Examples
Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/participant-01/video/interviews
Video genre	The genre contains classifications of the resource content: E.g. Folk song about local customs in Galicia	Short meaningful description, max. 10 words	Folk song about local customs in Galicia
Container format	The container format that contains the actual data in an encoded form, e.g. Matroska (MKV), Ogg, MP4	n/a	Matroska (MKV)

Media type	Standardised naming of the file format, a media type, e.g. video/mp4	n/a	video/mp4
Codec	Indication of the encoding method how the video/audio signals were converted for storage. E.g. mpeg-2	n/a	mpeg-2
Resolution info	The resolution of the video in pixels. E.g. 1280x720	Format: WIDTHxHIGHT, in Pixels, no spaces before and after the x; x is the separator between WIDTH and HEIGHT	1280x720
Frame rate (fps)	Number of frames per second.	Number of frames per second.	30
Duration	The duration of the resource in HH:MM:SS	HH:MM:SS	00:53:04
Duration of speech	The duration of the effective speech, the duration without interruptions, pauses etc. in HH:MM:SS	HH:MM:SS	00:40:10

6.1 Controlled Vocabulary

Field name EN	Controlled Vocabulary Values	Description
Container format	3GP	Third Generation Partnership, for lower bandwidth applications
Container format	ASF	Advances Systems Format

Container format	AVI	Audio Video Interleave
Container format	DVR-MS; WMV	Microsoft Digital Video Recording, the usual media type is video/x-ms-wmv
Container format	FLV; F4V	Flash video
Container format	MKV	Matroska Container
Container format	MJ2; MJP2	Motion JPEG 2000, the usual media type is video/mj2
Container format	QuickTime	Apple Quick Time Movie Container, this supports a broad list of codecs; please consult: https://en.wikipedia.org/wiki/QuickTime
Container format	MPEG-1/MPEG-2	MPEG Container; can contain MPEG-1 and MPEG-2 codec
Container format	MPEG-4 (MP4)	MPEG-4 Container, supported codecs are AVC (H.264), AV1, H.263, MPEG4 Part 2 Visual, VP9
Container format	Ogg	Multimedia container that can contain Theora, VP8, VP9 Codec.
Container format	WebM	Web Media container based on Matroska.
Container format	AVCHD	Can contain a H.264 video stream/codec.
Container format	RM	Real Media Container, the associated media type is normally application/vnd.rn-realmedia

Container format	other	n/a
Media type	video/mj2	Media type for Motion JPEG 2000 container; the usual file extensions are .mj2, .mjp2.
Media type	video/jpeg2000	Video sequence in jpeg 2000 images (used for streaming), the usual file extensions are .jp2, jpx and .j2k.
Media type	video/mp4	The associated container normally is MPEG-4 (MP4), this usually has file extension .mp4.
Media type	video/mpeg	The associated container is normally MPEG-1 or MPEG-2, the usual file extensions are .mpg, mpeg, or .mp1 resp. mp2.
Media type	video/x-flv	Media type for Flash Video, file extensions might be .flv, .f4v, .f4p, .f4a, .f4b
Media type	video/x-msvideo; video/avi; video/msvideo	The associated container is normally AVI, the file extension is .avi.
Media type	video/x-ms-wmv	Windows Media Files, the usual file extension is .wmv, the container is normally DVR-MS
Media type	video/x-ms-asf; application/x-ms-asf	Used for video streaming (e.g. video meeting platforms), the associated container is usually ASF.

Media type	video/3gpp; video/3gpp2; video/3gp2	Media type that is often used on mobile devices (e.g.mobile phones), the usual file extension is .3gp
Media type	video/x-matroska	The usual file extensions are .mkv, .mk3d, .mka, .mks
Media type	video/webm	Supported codecs are AV1, VP8, VP9, the associated container is WebM, the usual file extension is .webm
Media type	video/quicktime	The usual file extension is .mov, .qt
Media type	video/ogg	The usual file extensions are .ogg, .oga, .ogv, .ogx
Media type	application/vnd.rn-realmedia; video/x-pn-realvideo	Media type for real media videos
Media type	video/FFV1	
Media type	other	n/a
Codec	AV1	AO Media Video 1 Codec, possible containers are MP4, WebM
Codec	H.264	Advanced Video Coding, possible containers are 3GP, MP4. The default codec for MP4s.
Codec	H.263	H.263 video codec, possible containers are 3GP
Codec	H.265 (HEVC)	High Efficency Video Coding, possible container is MP4

Codec	MP4V-ES	MPEG-4 Video Elemental Stream, possible containers are 3GP, MP4
Codec	MPEG-1	MPEG-1 Part 2 Visual, possible containers are MPEG, QuickTime
Codec	MPEG-2	MPEG-2 Part 2 Visual, possible containers are MP4, MPEG, QuickTime
Codec	Theora	Theora, possible container is Ogg
Codec	VP8	Video Processor 8, possible containers are 3GP, Ogg, WebM
Codec	VP9	Video Processor 9, possible containers are MP4, Ogg, WebM
Codec	other	n/a
Frame rate	n/a	n/a

7. IMAGE METADATA

Field name EN	Description	Input rules	Examples

Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/media/images
Image genre	The genre contains classifications of the resource content: E.g. Photographs taken in Galician villages	Short meaningful description, max. 10 words	Travel
Type of image content	Description of the image content. E.g. Sign posts	Short meaningful description, max. 10 words	Sign posts
Compression	If the image is compressed or not compressed. E.g. Lossless compression	n/a	Lossy compression
Media type	Standardised naming of the file format, a media type. E.g. image/png	n/a	image/png
Raster or vector graphics	Indicates if images are raster or vector images. Typically, JPEG, GIF, BMP and TIFF are raster graphics, while SVG and AI (Adobe Illustrator) are vector graphic formats.	n/a	Raster
Resolution info	The resolution of the image in pixels. E.g. 890x720	Format: WIDTHxHEIGHT, in Pixels, no spaces before and after the x; x is the separator between WIDTH and HEIGHT	890x720

7.1 Controlled Vocabulary

Field name EN	Controlled Vocabulary Values	Description
Compression	Lossless compression	Lossless compression algorithms can restore the original data. Some file formats allow data stored with lossless or lossy compression. Lossless compression can be used with JPEG 2000, PNG, GIF and TIFF
Compression	Lossy compression	Lossy compression results in not restorable data loss and in smaller image sizes. Typically, JPEG uses lossy compression.
Media type	image/bmp	Bitmap Image Format, the usual file extension is .bmp
Media type	image/gif	Graphics Interchange Format, the usual file extension is .gif
Media type	image/jpeg	JPEG, the usual file extension are .jpg or jpeg
Media type	image/png	Portable Network Graphics, the usual file extension is .png
Media type	image/svg+xml	Scalable Vector Graphics, the usual file extension is .svg
Media type	image/tiff	Tagged Image File Format, the usual file extension are .tif or .tiff

Media type	image/jp2	JPEG 2000, the usual file extension is .jp2
Media type	image/dicom-rle	Digital Imaging and Communications in Medicine with Run Length Encoding Compression
Media type	application/dicom	Digital Imaging and Communications in Medicine Media type for DICOM structures with index of files (DICOMDIR)
Media type	application/dicom+json	Digital Imaging and Communications in Medicine Media type for bulk data including JSON to encode DICOM objects and image data and metadata
Media type	application/dicom+xml	Digital Imaging and Communications in Medicine Media type for bulk data including XML to encode DICOM objects and image data and metadata
Media type	other	n/a
Raster or vector graphics	Raster	Raster (Bitmap) if graphical information is stored in pixels that are arranged in raster.
Raster or vector graphics	Vector	Vector if graphical information is calculated with algorithms.

8. TOOLS

Field name EN	Description	Input rules	Examples

Reference path	Click on the directory tree icon to select a path to relate metadata to the whole corpus/data collection or to a subfolder of it.	n/a	/annotations
Software	The software used to generate, process, or annotate the data. E.g. EUDICO Linguistic Annotator (ELAN)	Please use the common designation of the software and the abbreviation in brackets: Software name (Abbr)	EUDICO Linguistic Annotator (ELAN)
Software role	The role of the software that was used for data annotation. E.g. audio capture, transcription	Please use a list of maximal five key words to describe the role of the software tool: Keyword 1, Keyword 2, Keyword 3, Keyword 4, Keyword 5	Audio capture, transcription
Description	Description of the Software as free text.	Free text	n/a