

CLARIN Standard Committee

Report of activities

www.clarin.eu/governance/standards-committee



What we do

The tasks of the CSC are:

- to collect, consolidate and prepare for publication in a single place its findings and recommendations related to standards;
- to maintain the set of standards supported by CLARIN and adapt them to new developments within or outside CLARIN;
- to publish and promote the standards supported by CLARIN;
- to develop and implement procedures for the discussion of recommendations and the adoption of new standards;
- to ensure harmonisation of standards between CLARIN ERIC and related initiatives;
- to ensure communication with international standards bodies such as ISO;
- to advise the BoD in all matters related to standards.

The current questions

What data standards are accepted for deposit by the various CLARIN B centers?





Can we create a coherent list of formats and MIME-types accepted?

What are the recommended formats for certain types of resources?

Item submission

- 1 Basic Info
- 2 Who's involved
- 3 Describe
- 4 Upload
- 5 License
- 6 Note
- 7 Review
- 8 Complete

Submission Info

 Corpus	 Lexical conceptual	 Language description	 Technology / Tool / Service
---	---	---	--

i Type of the resource: "Corpus" refers to text, speech and multimodal corpora. "Lexical Conceptual Resource" includes lexica, ontologies, dictionaries, word lists etc. "language Description" covers language models and grammars. "Technology / Tool / Service" is used for tools, systems, system components etc.

Title

i Enter the main title of the item in English.

Main objectives for the 2019-2020 cycle

- To build on the research originated by Dieter Van Uytvanck on the Key Performance Indicators (KPIs) relating to the percentage of Centres that publish explicit information on what data formats they accept.
- From that, we will be able to see what the most common formats are as currently recommended in the bottom-up fashion, by the individual centres.
- Not all centres have answered our requests yet, but we are nearly there, and the current state of this research can be seen at https://docs.google.com/spreadsheets/d/1icWkilaVxFW08NB3qXPY_PGChxoUqsg9-v5xZ86Ezmo/edit?usp=sharing
- We have also formulated certain points of advice that we would be happy to offer to the centres interested in improving the CLARIN KPI by explicitly publishing their information.

Our current activity



Formats and Mimetypes



File Edit View Insert Format Data Tools Add-ons Help



Category	A	B	C	D	E	F	G	H	I	J	K	L	M
Category	Name	Mimetype	Extension	TLA	Cocoon	UDS	HZSK	IDS	BAS	ARCHE	IRTOLAN	IMS	
3D	WaveFront Object	text/plain	.obj										
3D	Polygon file format	text/plain	.ply										
3D	X3D	model/x3d+xml	.x3d										
3D	COLLADA	model/vnd.collada+xml	.dae										
Audio	Wave	audio/x-wav	.wav	1	1	1	1	1		1	1		
Audio	FLAC	audio/flac	.flac		1			1		1	1		
Audio	AIFF	audio/x-aiff	.aiff					1	1		1		
Audio	PhonDat 1	??	??										
Audio	PhonDat 2	??	??										
Audio	Ogg Vorbis	application/ogg	.ogg			1						1	
Audio	MPEG 4 audio	audio/mp4	.m4a									1	
Audio	MP3	audio/mpeg	.mp3			1						1	
Audio	RAW	audio/raw	.raw						1		1		
Audio	NIST SPHERE	audio/x-nist	.nist						1				
Audio	Matroska	audio/x-matroska	.mka										
Audio	BWF	audio/x-wav	.bwf										
Audio	MXF	application/mxf	.mxf										
Audio	OPUS	audio/ogg; audio/opus	.opus									1	
Compression	GZIP	application/gzip	.gz			1							
Compression	ZIP	application/zip	.zip			1							
Compression	RAR	application/x-rar	.rar										
Computer Aided De	AutoCAD DXF version F	application/dxf	.dxf										
Computer Aided De	SVG	image/svg+xml	.svg										
Computer Assisted	REFI-QDA												
Container	TAR	application/x-tar	.tar										
Databases	CSV	text/csv	.csv							1	1		

Objectives

All centres should publish a list of recommended formats for deposit (also in English)

Produce an organised list, to be used as a template.

Update the list of relevant standards

Relevant standards

The table below contains a (non-exhaustive) list of standards that are relevant for the CLARIN community.

Source: CLARIN standard guidance website (provided by the IDS)

Abbreviation/Name	Topic(s)	Standard body	CLARIN centre(s)
CES	Generic Corpus Annotation	EAGLES	
CHAT	File Formats Transcription	Other	
CMDI	Metadata	ISO	CLARIN-PL, IDS, MPI-PL, UC
Controlled Vocabulary	Controlled Vocabulary Knowledge Representation Thesaurus	NISO	CLARIN-PL, MPI-PL
CQLF	Query	ISO	IDS
DCAM	Metadata	DCMI	
DCMES	Metadata	DCMI	MPI-PL
DCR	Data Categorization	ISO	CLARIN-PL, IDS, MPI-PL

<https://www.clarin.eu/content/standards-and-formats>

(Partial) recommendations

- include a timestamp/version number in the list
- use the English language (in addition to the member language)
- distinguish between recommended and "others", to stress (and reflect) what centres want, not what they have to accept because users bring these formats.
- use an appropriate degree of detail (XML is usually not specific enough) and recommend best practice format parameters (e.g. plain text in UTF-8 where possible)
- Follow the format of one of the already existing lists

Questions?

Would you like to know how to add info on your centre?

Would you like to get help to add the relevant info on your centre's webpage?

Would you like to help?

Please come to our Bazaar stand on Wednesday!