Pursuing the elusive KPI: Filling the gaps in centre self-published standards-related information

CLARIN Standards Committee

What we do

The tasks of the CSC are (quoting from the 2019 Bylaws):

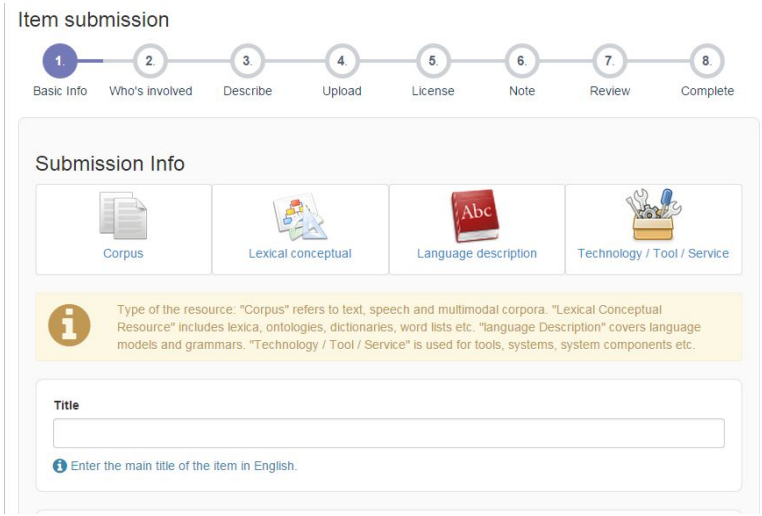
- to collect, consolidate and prepare for publication in a single place its findings and recommendations related to standards;
- to maintain the set of standards supported by CLARIN and adapt them to new developments within or outside CLARIN;
- to publish and promote the standards supported by CLARIN;
- 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 (but not limited to) 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?

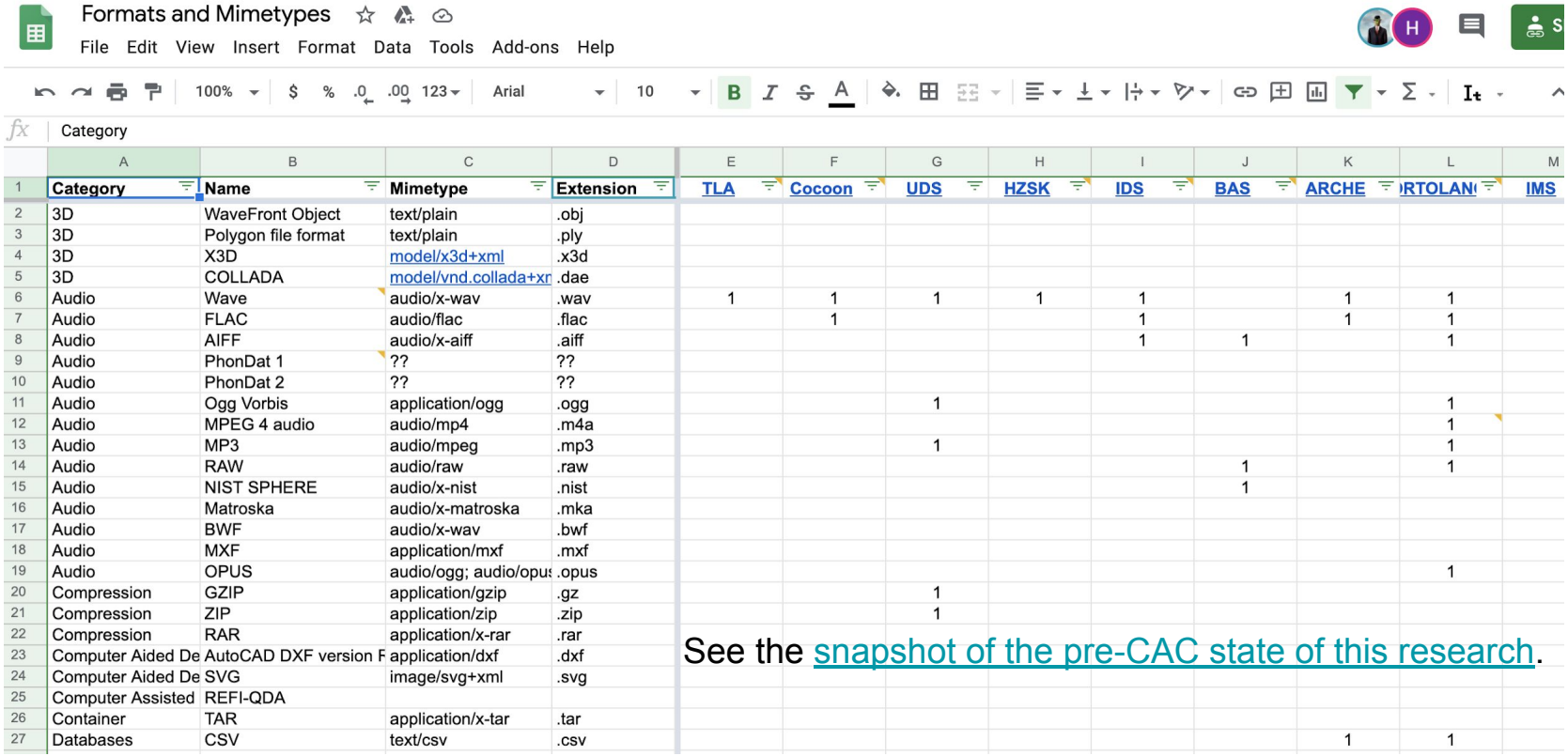
What are the recommended formats for certain types of resources?



Main objectives for the 2019-2020 cycle

- To build on the research originated by Dieter Van Uytvanck for the BoD 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 published explicit lists of formats
    - Many use very general statements “XML or any machine-readable format”
    - Some point to obsolete CLARIN-related standards lists of various kinds
    - Please see the section at https://www.clarin.eu/content/standards-and-formats#formats
    - If your centre is missing, please consider changing that!
    - (We will be happy to offer some hints, too)

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
1	Category	Name	Mimetype	Extension	TLA	Cocoon	UDS	HZSK	IDS	BAS	ARCHE	RTOLAN(	IMS
2	3D	WaveFront Object	text/plain	.obj									
3	3D	Polygon file format	text/plain	.ply									
4	3D	X3D	model/x3d+xml	.x3d									
5	3D	COLLADA	model/vnd.collada+xr	.dae									
6	Audio	Wave	audio/x-wav	.wav	1	1	1	1	1		1	1	
7	Audio	FLAC	audio/flac	.flac		1			1		1	1	
8	Audio	AIFF	audio/x-aiff	.aiff					1	1		1	
9	Audio	PhonDat 1	??	??									
10	Audio	PhonDat 2	??	??									
11	Audio	Ogg Vorbis	application/ogg	.ogg			1					1	
12	Audio	MPEG 4 audio	audio/mp4	.m4a								1	
13	Audio	MP3	audio/mpeg	.mp3			1					1	
14	Audio	RAW	audio/raw	.raw						1		1	
15	Audio	NIST SPHERE	audio/x-nist	.nist						1			
16	Audio	Matroska	audio/x-matroska	.mka									
17	Audio	BWF	audio/x-wav	.bwf									
18	Audio	MXF	application/mxf	.mxf									
19	Audio	OPUS	audio/ogg; audio/opus	.opus								1	
20	Compression	GZIP	application/gzip	.gz			1						
21	Compression	ZIP	application/zip	.zip			1						
22	Compression	RAR	application/x-rar	.rar									
23	Computer Aided De	AutoCAD DXF version F	application/dxf	.dxf									
24	Computer Aided De	SVG	image/svg+xml	.svg									
25	Computer Assisted	REFI-QDA											
26	Container	TAR	application/x-tar	.tar									
27	Databases	CSV	text/csv	.csv							1	1	

See the snapshot of the pre-CAC state of this research.

(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 (see next slide)

Examples to follow

- Bavarian Archive for Speech Signals
- The Language Archive
- Data Archiving and Networked Services
- Tübingen Archive of Language Resources
- A Resource Centre for Humanities Related Research in Austria
- ….

Thanks!

















https://www.clarin.eu/governance/standards-committee