CLARIN:EL Recommended File Formats

Guidance on selecting file formats for long-term accessibility and interoperability

This page lists the file formats which are recommended for depositing in CLARIN:EL. If you have a suggested update or a question, the CLARIN:EL technical helpdesk (technical-helpdesk@clarin.gr) will be delighted to hear from you.

File Formats for Digital Preservation Policy

To ensure access and usability of your data to the broadest audience into the long term, the CLARIN:EL team has considered the following factors to determine which file formats are recommended in CLARIN:EL infrastructure:

- Processability:
    - Suitability for the type of resource and/or type of processing.
    - In order to be processable by the CLARIN:EL integrated NLP workflows , textual data have to be in one of the formats that the workflows can process (listed below)
- Preservation
    - Suitability for research by the designated communities.
    - How widespread the format is: broadly used formats, not deprecated, known to the designated communities.
    - Use of open source rather than proprietary format.
    - Whether the format employs lossy or lossless compression.

The policy, which is based upon the above-mentioned factors, meets the mission of CLARIN:EL to collect, preserve and distribute digital language resources and language processing services for the support of researchers, academics, students, language professionals, citizen scientists and the general public. In order to arrive at the appropriate recommendations for individual file formats, or to decide on their suitability for particular

kinds of research activities/types, the purpose for which they are intended has to be considered. For example, while PDF/A has been developed for unproblematic long-term archiving and is an excellent format choice for documentation, it is undoubtedly *not* suitable for textual data intended for language processing. Therefore, based on the types of resources that are in the scope of the CLARIN:EL user communities and the processes offered/supported, the CLARIN:EL team discerns the following set, pertinent to the field of digital language resources, for which specific recommendations are provided:

- **CLARIN:EL processable data:** Textual data that can be input data for CLARIN:EL workflows
- **Textual Data**: Written unstructured/plain text or originally structured text (e.g., HTML) without linguistic or other mark-up added for research purposes (non-processable by the CLARIN:EL workflows)
- **Text Annotation**: Annotations of textual source language data, with the original text included or as a stand-off document
- **Language Description**: Data that describe a language or some aspect(s) of a language via a systematic documentation of linguistic structures (Grammars, Machine learning (ML) models, N-gram models)
- **Lexical / Conceptual Resource**: A resource organised on the basis of lexical or conceptual entries (lexical items, terms, concepts etc.) with their supplementary information (e.g., morphological, semantic, statistical information, etc.)
- **Image data**: Digitized images of analogue sources of written language data for research purposes (e.g., scans of handwriting, photos of inscriptions) or two-dimensional pictures or figures that are distributed with associated textual data for NLP analysis (e.g., medical images (image data) accompanied with radiological reports (textual data))
- **Audio data**: Audio recordings providing spoken language data for research purposes (e.g., audio files with the pronunciation of words for a lexicon, recorded interviews, radio broadcasts, etc.)
- **Video data**: Video recordings providing multimodal or sign language data for research purposes.

Format Recommendations

Formats that fulfil the criteria of the Digital Preservation Policy, mentioned above, are preferred; however, additional formats are accepted, as a "first-entry level", with the proposal for conversion to recommended formats.

Therefore, file formats are categorized into two preservation levels (recommended, acceptable) always in the context of each case. The acceptable list is not exhaustive, especially in the case of text annotation, but rather indicative, and it is proposed for an acceptable format to be converted to a recommended format.

	Recommended	**Acceptable**
CLARIN:EL processable data	**Monolingual textual data**: plain text **Monolingual encoded data**: XCES-ILSP variant (XML based format compliant with the XCES model for corpora) **Bi-/Multilingual encoded data**: TMX (XML based format for aligned data), MOSES (text-based format for parallel data)	
Textual Data	**File Formats**: plain text **Formatted/Encoded**: ODT, DOCX, PDF/A, HTML, Latex, TeX, MOSES	PDF, SGML, Rich Text Format (.rtf), Microsoft Word (.doc, .docx), PostScript

Text Annotation	**File Formats**: XML, XMI, CSV, TSV, RDF (all serialisation formats RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), JSON **Models**: XCES for corpora and structural annotation, TEI for structural and linguistic annotation, GrAF linguistic annotation, TMX for aligned, GATE linguistic annotation, CoNLL family (CoNLL-U, CoNLL-2000, CoNLL-2002, CoNLL-2003, CoNLL-2006, CoNLL-2008, CoNLL-2009, CoNLL-2012) for linguistic annotation, NIF linguistic annotation for RDF data, WARC for web crawled data	**File Formats**: SGML, Plain Text, Microsoft Excel (.xlsx, .xls), ELLOGON
Language Description	**ML Model**: H5, ProtoBuf, ONNX, PMML, Pickle, MLeap, YAML, JSON **N-gram model**: ARPA	
Lexical / Conceptual Resource	**File Formats**: XML, CSV, TSV, RDF (RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), OWL **Models**: LMF for lexica, OWL for ontologies, SKOS for thesauri, OntoLex-Lemon for lexica, TBX for terminological data	Microsoft Excel (.xlsx, .xls), Plain Text, SQL

Image data	**All images**: TIFF, SVG, JPEG 2000, PNG, GIF **Scanned images**: PDF/A	JPEG, BMP, Photoshop, NifTi, FlashPix, PDF
Audio data	WAV, AIFF, FLAC	MP3, MPEG, Windows Media Audio
Video data	AVI	MPEG-4, RealNetworks 'Real Video', Windows Media Video, Flash Video, QuickTime Video