DOBES Format and Encoding Agreements

This guide describes the DOBES Agreements regarding archival format and encoding standards as they were defined in the pilot phase and emerged slowly due to technological advancements.

General Framework:

- The DOBES Language Archive has to serve two main purposes: (1) to give direct access to the documentation data where access is permitted and (2) to preserve the material for future generations.
- The DOBES programme is a project comprising many documentation teams that deliver (copies of) all material to the DOBES archive in order to create a coherent repository of language resources.

1. General Remarks

We have to distinguish between **archival formats and encodings** (F+E) and those that are offered by the donators. Archival F+E have to adhere to open standards, while F+E offered by the donators have relate to workflow agreements between donators and archivists. They depend on the kind of manipulations and conversions the archivist is able to handle. Naturally, the set of offered F+E is a superset of the archival F+E.

We also have to distinguish between **representation** and **presentation** F+E. Representation F+E are relevant for the neutral and persistent storage of data, while presentation F+E determine how data is presented to the users. In this note we will mainly be concerned with representation formats, since presentation F+E can be derived from the back-end ones.

In general, **encodings** describe how information such as characters or video streams are represented as bit-patterns. **Formats** describe how these patterns are packaged into archival objects, i.e. files.

Since technology is changing and therefore the agreements will also change over time, we have to be explicit about the date this document was created: **April 2005**

2. Archival Encodings

The following encodings are accepted for the archive:

- For character encoding UNICODE is the standard, i.e. all input material has to be transformed to UNICODE and people have to create UNICODE where possible.
- For cataloguing, the IMDI standard is obligatory (XML plus IMDI schema).
- Structured textual material (annotations and lexica) has to be structured with XML and they have to adhere to generic schemas. At this moment the DOBES archive uses EAF (ELAN Annotation Format) and LMF (ISO Lexical Markup Framework) as underlying schemas.
- For unstructured texts the archive will accept plain text and HTML. In exceptional cases PDF documents have to be accepted as well.
- For digital audio representations, linear PCM is the standard. In general 16 bit (96 dB) and 44.1/48 kHz are sufficient, however, 24 bit (144 dB) and 96 kHz are also accepted. (MiniDisc and MP3 are not accepted as archival back-end encodings!)
- For digital images, JPEG, PNG and TIFF are the accepted standards. In the future we expect increasingly often RAW/TIFF as representation formats, since JPEG implies lossy compression.
- For digital video, MPEG2 is the current archival standard. Some old resources may be in MPEG1 which has to be accepted. Audio information included in the video streams should be extracted as linear PCM data as well.
- For video, both PAL and NTSC norms are accepted.

Since we know that conversions are often not free from information loss, we will also store some original files in the archive such as for Shoebox. For presentation purposes, we will generate MP3 and MPEG4 files from the above-mentioned back-end formats.

DOBES Format and Encoding Agreements 2

3. Archival Formats

For some encodings, the formats are predefined by the chosen encoding standard.
The following formats are accepted for the archive:

- For structured texts, validated XML files are accepted (see above).
- For unstructured texts plain text or HTML files are accepted. Exceptions in PDF are possible.
- For audio streams, WAV file packaging is the standard.
- For images JPEG, PNG or TIFF file packaging is accepted.
- For videos, all formats are converted to MPG formatted files.

4. Input F+E

This chapter may be subject of changes, i.e. no one can rely on the availability of the mentioned services. In many cases conversion is not trivial, i.e. it is strongly recommended that early workflow agreements are made. In addition to the above mentioned F+E, the following F+E are accepted as donator F+E:

- Shoebox/Toolbox annotation, lexicon and auxiliary files such as type file etc; we recommend the use of Toolbox to deliver Unicode
- Transcriber annotation files
- CHAT annotation and lexicon files with the recommendation to use Unicode encoding
- MiniDisc recordings (although not recommended due to severe filtering)
- DV as native video encoding created by most of the current cameras
- a number of media formats such as AVI, MOV, MPV, MPA

In the past, the DOBES archive also processed other F+E, but this has to be negotiated between teams and the archivist, since some of the conversions included are very time consuming.

- Carrier norms: MiniDisc, Cassette, Reel-to-Reel, Uher4400, Hi-8, UMatic, VCD
- Document Types: WORD files with clear and consistent structure marking
- XML: other XML files with consistent tagging even if no schema is provided
- Character Encoding/Fonts: SIL IPA Sets, IPA Kiel

5. Recommended Tools

Tools are not part of the agreements, although we can give a number of recommendations for tools from which we know that the agreed F+E can be generated. This list cannot be seen as complete, since there may be more tools that support the mentioned F+E. In particular, it does not make sense to list all useful media tools.

- Shoebox/Toolbox: to create Shoebox formats for lexica and annotations, Toolbox is highly recommended, since it allows for working with Unicode
- Transcriber: to create Transcriber files
- CLAN Tools: to create Chat formatted files; we recommend the use of the Unicode version
- ELAN: to create EAF files and to import/export Shoebox, Chat and Transcriber annotation files
- Onto-ELAN: to create EAF files making use of ontologies
- IMDI Editor: to create validating IMDI files
- LEXUS: to create LMF files and to import/export Shoebox and Chat lexicon files

Often in this text we speak about UNICODE. Mostly, the flexible UTF-8 encoding scheme is applied by the software.