Technology and Tools for Language Documentation

Peter Wittenburg, Romuald Skiba, Paul Trilsbeek
MPI, N megen

Language documentation and preservation
When speaking about technology and tools for language documentation, there are two important aspects to keep in mind. First, technology and tools are continuously changing. Something that is state of the art today can be outdated tomorrow. Second, people involved in language documentation do not all have the same needs and preferences, and therefore may have diferent criteria for making choices. Documentation creators, for example, are typically concerned with familiarity and ease of use, whereas for archivists, technical quality and long term preservation are more important.

The language documentation process consists of several steps (some of these take place in sequence, others in parallel):

- creation of recordings
- transferring and manipulating recordings using computers
- transcribing, annotating and pre-analysing recordings
- integrating materials into an archive infrastructure
- exploring and re-analysing materials for various purposes
- making materials accessible to diferent user groups
- protecting materials against misuse
- preserving materials for use by future generations

For each of these activities we can identify relevant methods, standards and frameworks, the sum of which we may call technology. Technology can be applied through the use of tools - e.g. DV technology is used in digital video cameras and XML technology is used in the annotation tools Transcriber and ELAN. Language archives need technology and tools to support the ongoing management of data and to provide for various usages of data by diferent user groups.

How can we satisfy diferent user groups?
We envisage that the following groups of users may wish to access endangered languages archives, each group bringing their own specifc needs:

- language communities and linguists who would like to access language material for educational or related purposes may require an educational style of presentation.
- local centres may want to have complete copies of digital archive materials, including metadata and structural information, in order to provide local, fexible usage of the data.
- linguists and other researchers may want to access a cross-linguistic selection of data for a typological study.
- teachers of linguistics may want to demonstrate to students how languages can difer by comparing annotated recordings of diferent languages.
- journalists may want to create a story about language diversity for the interested public in collaboration with linguists or members of a speech community.

While archives might aim to support a large variety of usages, it is almost impossible to create customised access tools for every conceivable group of users. What archives can do is provide access to data in such a way that users can easily fnd and retrieve the data they need. Therefore, an archive can ofer:

- well defned and well documented data formats - preferably open
- extensible archive and data structures
- detailed metadata descriptions
- powerful search and exploration tools for content and metadata
- easy, yet secure, access to the data

Users can use search and exploration tools to create their own resources from the archive. Or they can use information about the archive and its data structures to help create their own access and presentation tools for their own local purposes.

Media, codecs, formats—can we predict the future?
Today's language archives almost exclusively archive material in digital form. This has advantages such as perfect reproducibility; however, there are also great issues to deal with such as the limited lifetime of storage media and the continual changes in data formats. To ensure that materials can still be read after some hundreds of years, an archive has to make the migration

of data to the latest storage media, file formats, and codecs as easy as possible.

In traditional archives, materials can be accessed using the human senses (typically, the eyes). Although it may be necessary to understand a special Sumerian encoding system to understand the content of a clay tablet, the eyes alone can identify signs and their patterns. In digital archives, material is stored as magnetic sequences of ones and zeros, so we always depend on computer hardware and software to provide access to the content.

Even after the magnetically encoded data has been accessed, it may not be directly usable. We have to recognise various layers of encoding, such as are required for audio and video:

- codecs that determine how audio and video streams are represented by bit streams (a parallel for text would be how particular glyphs are encoded by particular sequences of bits)
- file formats that determine how these bit streams are packaged into units that are the objects handled by operating systems and application software
- tools that process bit streams or present them on screen or in print

Examples of video codecs are MPEG1, MPEG2, MPEG4 and DV, while video file formats include AVI and MPG. Examples of audio codecs are Linear PCM, MP3 and MD-Atrac, while common audio file formats are WAV and AIFF. This layered system means that an AVI file can contain video streams with different codecs, so the specifcation "AVI file" does not identify the encoding or quality of its video content. Various software tools support different codecs and formats, so for each tool one has to check what is supported.

For archiving purposes, open and well-documented standards such as MPEG, UNICODE and XML should always be used. It can be assumed, for example, that MPEG codecs will continue to be used for many years and that there will therefore be tools available that can decode MPEG bit streams. For encoding characters, UNICODE is recommended despite current limitations in the range of characters represented. For structuring documents, XML is recommended; in addition to being a widely adopted standard, XML files are human readable, and can be viewed and edited using even the simplest text editor.

Software tools that encapsulate information content in a proprietary format - such as MS Access, FileMaker Pro, MS Word and Excel - do not provide files that are appropriate for archiving. There is a contradiction between the short-term needs of linguists and the long-term needs of archivists. While linguists prefer tools that efciently provide data entry and presentation, archivists are more concerned with data representation, i.e. encoding and format standards.

For continuously time-varying signals, such as audio and video, archives prefer to store the most simple, direct and high quality digital representations of signals. For audio, a procedure called Linear PCM defnes a temporal resolution (such as 44.1 or 48 kHz), measures the value of the signal's amplitude in a particular resolution (e.g. 16 bits) at equidistant points in time, and then stores those values in sequence. The amount of data produced in this way is not so great as to cause problems for storage; compression such as MP3 or MD-Atrac is not necessary. However, many recording devices do apply these compression techniques, which remove spectral and temporal components from the original signal (components which are claimed to be fltered out by the human ear anyway). Due to this unrecoverable reduction in information, and the additional complexity in decoding compressed files, it is generally recommended not to use compressed formats for material to be archived.

Turning to images, all digital still cameras produce files in the JPEG file format which has become the de facto standard even though it applies lossy compression which deletes high-frequency image components. However, many cameras are now ofering RAW file format, which is not only uncompressed but allows adjustments of settings (white balance, ISO sensitivity, etc.) after shooting. RAW files are of course larger in size, but with fash memory prices dropping rapidly, this is no longer a problem.

For video, however, current technologies are not able to handle uncompressed video streams. Various compression formats such as MPEG1/2/4 and DV have been developed. Each of these has its own disadvantages. DV is used by almost all camcorders, but its data rate is too high for current storage media. MPEG2 can currently be seen as a good compromise between archiving needs and tractable data rates, while MPEG1 and MPEG4 can be seen as derived formats for special purposes.

The choice of tools and technologies infuences the quality, stability and durability of materials. While not all the language documentation steps listed at the beginning of this article are directly related to archiving, the tools and technologies used in all of those steps ultimately afect the quality and stability of the archive that holds the materials. Therefore, to characterise or evaluate an archive, documentation of the technologies and tools used in all phases of material creation is needed.