Language Archive
Newsletter

Editorial

Dear members of the DOBES programme,

many of you were heavily involved in preparing the summer school and the conference which meant much extra work in addition to your documentation work. Due to these activities we tried to keep the number of emails and requests to you at a low level. Of course, we hope that both – summer school and conference in 2004 – will become a great success. The DOBES programme is very well-known and has strongly influenced the work in the endangered languages documentation domain and we are sure that both - conference and summer school - will add to the good image of the work in DOBES.

Now in the beginning of 2004 we would like to take up some points again that are relevant from the archivist's point of view. Some of the points have to do with the fact that the first teams will finish their work in the coming year. Others are just to inform you.

Given that we will not have so many exchanges as in the pilot phase where we had three intensive workshops within almost a years time, it seems to be necessary to exchange relevant information in a more regular way. Therefore, we thought that it would make sense to regularly circulate an electronic newsletter. It will also contain detailed information which may not be relevant for everyone. So you can pick out that information which seems to be relevant for you at a given moment. Each issue should also have space for one major contribution of wide interest. The newsletter will be archived and available at the DOBES web-site. We intend to distribute the Language Archiving Newsletter four times per year and will work to improve the layout. The next deadline for contributions is March 31st, 2004.

This issue is almost completely filled with items that come from the archiving team, but this should definitively not be the case in future. In contrast we would like to motivate you to participate with information that is relevant for many of us. We will also try to motivate others from the DELAMAN network to participate and send contributions. Currently, the three of us are responsible for the content, but we can imagine that other colleagues from the DOBES programme or even from outside join the editing group. All comments that can improve LAN are welcome.

Please, send all comments, questions and contributions to the following email address: LAN@mpi.nl

We wish you all the best for your work in 2004.

Hennie, Roman, Peter

Report Archiving Team

The archiving team created a report about its work mainly in 2003 that is available now on the internal DOBES web-site. It contains also information about the state of the archive as far as we can see it. Some of its points will also be mentioned briefly in this LAN issue. If you are interested in more details, please, look at the report.

Points for finishing Projects

Knowing that a number of teams in DOBES will finish their work in 2004, it would be important to check the status of the data exchange early enough. The archiving team needs to get information from you about the following points and will send you separate emails:
Which data (in particular annotations and other textual material) will be sent in which formats at what time?
Are all the tags that are used well-documented and described?
Is the current archive content correct and is your archive structure (metadata tree) correct?
Do all the metadata descriptions contain the correct information?
We would like to ask your help in answering these questions as soon as possible, since we expect a peak load at a certain moment to finish our part of the work.

Training Course

There will be another training course especially for new teams. But there is no problem to let also a number of "old" DOBES team members participate. The training course will take place in Nijmegen from May 10 to 14 . We will revise the program slightly such that more experienced users can drop in later. As in all the

other years there should be a session to discuss documentation issues. If there are members from the current documentation teams who would be interested to contribute we would be very happy. Please, let us know whether you have time and whether you see relevant topics to be transferred to the new teams that cannot wait until the big September events.

Access Rights Management
We are developing a central system to manage access rights for the resources that are contained in the archive at the MPI. The report gives more detail. The development work of the current version was finished in January 2004 and we are now testing it with MPI researchers very carefully.

It allows to define a hierarchy of access rights managers to share the amount of work and give the DOBES team leaders the opportunity to define read rights for their collaborators, members of the communities etc. There will be a manual as well and we will optimize the integration of the new web-pages after the work is finished. As soon as the tests are succesfull we will inform you how to use this system.

Access Methods
As some already may know we have changed the focus of our work slightly from technologies that allow us to create a stable and maintainable corpus to technologies that allow accessing the resource collection in an easier way. Therefore, a few new options have been developed and we are still testing and improving.

First, it is possible to navigate in the IMDI domain with normal web-browsers. The pages also offer a full-text search field that works on all metadata entries including descriptions. You can access these methods by using "http://corpus1.mpi.nl/BC/IMDI-corpora/".

Second, we are testing ways to offer the resources themselves via normal browsers. Solutions can be found in presenting the textual content with html such that media fragments can be offered by clicking on annotations or by using SMIL – a W3C standard – for media streaming with synchronized subtitles. In 2004 we will offer such solutions and expect some interaction with the DOBES members.

IMDI Developments
The new IMDI version 3.02 is finished, the tools were adapted and currently the manuals are being adapted. We will switch to the new version in February, i.e. all IMDI files will be transformed automatically. We will send an email, distribute the manual and would like to ask you then to follow the new guidelines. Contributions from DOBES members again were very important to create this new version that contains some simplifications in particular to describe the content.

All information and the tools are already available at (www.mpi.nl/tools or www.mpi.nl/IMDI). Since the semantics of the content description elements are changing, we will store the old IMDI descriptions.

ELAN Version
A new ELAN version with many new features was finished and is available on the web (www.mpi.nl/tools). The new version has a completely new user interface with improved ergonomics. More details are mentioned in the report and of course the manual that is currently being written will give all details. It should be mentioned again that there is no duty to use ELAN, it is just an option where you can expect support from us. The archive will rely on the EAF (ELAN Annotation Format) that is based on an open XML schema, i.e. we will convert all annotations to the EAF format. Only when there is a widely accepted standard we will switch.

Lexicon Component
Due to all the other work items we delayed the start with the concrete work on a lexicon component. However, in the new year we will finish our design work and start implementing such a component. The SHAWEL tool is seen as a design study and the good ideas will be re-used. SHOEBOX is still seen as an excellent example and whatever we will develop it has to be able to exchange data with SHOEBOX. Not all features of SHOEBOX can be implemented in a short time frame. The intention is to have some useful functionality ready at the end of 2004.

New Web-Site & Demo
As you will know we have created the demonstration DVD for the Science+Fiction exhibition with your help. In the mean time this was extended and also converted to an English version, since the demo will also be shown in cities such as Stockholm, Lissabon etc. Currently, we are busy to adapt this demo to make it the DOBES web-site and replace the old one. This work is finished now. We would like to ask for your help in extending it by contributions from your teams. A template for team contributions was developped and filled in for the Trumai team such that you can look at it. We will send a separate email about this issue when the demo can be viewed via the web. It will become available in January and we will send each team also a demo DVD.

Music Annotation
Some discussions took place to discuss the special issues of ethnomusic annotations within DOBES. Sven Grawunder took over the function of coordinating this work. At this moment there are many unclear aspects, of course. Sven and Peter will write a small report which will be sent first to the group of people interested in this topic. There are excellent tools around for music annotation and it has to be checked what is missing and how we can extend ELAN for music-ethnology work without re-inventing the wheel.

Ethnology in Language Documentation
Bruna Franchetto and Martin Gänszle will co-ordinate the work about requirements from the ethnologists within DOBES. They will create a list of interested persons and exchange documents.

Long-Term Archiving
Some energy was invested to tackle the issue of long-term survival of data. There is no perfect solution but it is our task to increase the chances that DOBES data (and others about endangered

languages) will survive enabling future generations to make use of it. It is clear that the vulnerability and the limited lifetime of our current magnetic and optical storage media can only be dealt with by a continuous migration to new media and by distributing the data to several archives.

The Archiving Team has taken the following steps
(1) At the MPI there are 3 automatically and dynamically generated copies of the data, one in a different building.
2) The DOBES data is now dynamically stored at the GWDG computer center in Göttingen as 4th copy. A 5th copy is being prepared to be dynamically exchanged with the other big computer center of the MPG in Garching. Both computer centers themselves have strategies for redundant copies, i.e. actually there are even more copies of the data.
3) Discussions have been started to back up these technological solutions by institutional guarantees.
4) In the realm of the DELAMAN initiative it is clear that serveral archives are interested to exchange their holdings such that even more copies are created and distributed world-wide. This world-wide distribution can hopefully be realized within the coming 5 years. It should be mentioned that all our long-term efforts are focusing on bitstream preservation. We don't believe that there is a simple and inexpensive way to solve the long-term interpretation problem except to rely on the intelligence of future generations.

DELAMAN Initiative
The DELAMAN network (Digital Endangered Language and Music Archives Network) was founded by the following archives: AILLA (Austin), DOBES (Nijmegen), E-MELD (Ypsilanti), HRELP (London), PARADISEC (Sydney). A number of topics will be tackled within this collaborative framework which can be seen at the preliminary web-site (www.delaman.org). Two meetings were organized until now to discuss the goals, the organizational basis and the topics to work on.

Forschungsprofessur Himmelmann
Nikolaus Himmelmann (Waima'a Project, Bochum) has been granted a Forschungsprofessur by VWS to work on prosody in language documentation within the DoBeS framework. The analysis and representation of prosodic units, in particular prosodic and intonational phrases, is one of the least developed areas in language documentation and description. This reflects the fact that prosodic analyses have played a somewhat marginal role in general linguistics for quite some time. However, in the last two decades the analysis of prosodic phenomena has made considerable progress in a number of regards and theoretical frameworks, including prosodic phonology (syllable and word prosodies), intonational phonology (tonal analysis of intonation contours) and discourse-based grammar (the role of prosodic chunking and prosodic features in verbal interactions).

The basic goal of the project is to make available insights from these different developments for purposes of language documentation, i.e. to provide for means which increase the overall quality of a language documentation by paying explicit attention to prosodic phenomena.

More detailed info will follow once the project is under way (currently the starting date is 1 April 2004). But other DoBeS teams who are interested to get help or advice on prosodic problems (in particular intonation/representing intonation units in their transcripts) or are willing to share interesting prosodic data are welcome to get in touch with Nikolaus at himmelma@linguistics.ruhr-uni-bochum.de.

Special Contribution

Digital Formats for Images, Audio and Video
Peter Wittenburg, Reiner Dirksmeyer, Hennie Brugman, Gerd Klaas (MPI for Psycholinguistics)

Digital audio and video formats became very important when the transition was made from traditional magnetic tapes as the main recording and preservation technology to computer-based digital methods. Increasingly often linguists and other collaborating researchers are confronted with acronyms such as MPEGx, MD, AVI etc but don't exactly understand where these are about and what implications decisions may have. This paper wants to give a brief overview about the most relevant encoding formats (codecs), file formats and application programming interfaces (APIs). This may help interested users to find their way in the jungle of acronyms.

We have to distinguish three levels: (1) The stream of data – be it the acoustic wave – has to be encoded digitally such that it can be represented and processed as sequences of "0" and "1" on our digital machinery. We call this the encoding of media streams. At the playing side consequently their has to be a decoding process that creates waves or moving images again. Therefore the term "codec" (coder/decoder) is often used to refer to this level. (2) The stream of data has to be packaged into file formats that can be handled by the operating system and application programs. Application programs need to know how to interpret the chunks of information contained in a file and how to retrieve typical metadata such as sample frequency. This information is normally contained in headers (first x bytes in a file) in a specific sequence and format. (3) Some acronyms used in the media world refer to more complex objects (media and linked annotations). To write and read such complex objects the designers as in the case of Quicktime MOV files offer APIs (Application Programmer Interfaces), i.e. to access such files one has to use the provided programs or develop a new one making use of that API.

Audio Encoding Standards
Due to technological developments a number of different standards emerged for the encoding of acustic waves. **Linear PCM**[12] (Pulse Code Modulation) is the most simple and direct way

---

[1] For certain telephone and other applications codecs were developed that use non-linear methods in both dimensions: voltage and time. These (ADPCM, A-law, μ-law, ...) will not be commented here.

to digitize a waveform. The given voltage range is divided in equidistant steps. For 16 bit processing this means that waveforms can be represented in more than 64.000 equidistant values. In general this is said to be sufficient for speech signals. For signals with a large dynamic range – where one has very silent and very loud parts as for example in Beethoven symphonies but still wants to preserve the details – some archives recommend to use 24 bits. We assume that for the work within DOBES 16 bit is sufficient. Each value is represented by 2 bytes. In general the samples of the waveform are taken at equidistant time steps defined by the sampling frequency. According to Nyquist only those frequency components are represented for which at least two points are taken for one period. Speech – in particular from children and female voices – are said to have frequency components to up to 7 kHz. Therefore, for speech recordings 16 kHz is the minimal sample frequency to cover the details. For Hifi recordings signal components of up to 20 kHz are seen as relevant[3], therefore the sample frequencies for most state-of-the-art equipment is set to 44.1 or 48 kHz. Some argue that over-sampling should be done and therefore require 96 kHz for archive recordings. We assume that 44.1 or 48 kHz is sufficient for all work in DOBES.

Both the **ATRAC** (used in **MiniDisc** Recorders) and the **MP3** (used in stand-alone recorders and in the MPEG video codecs) compression algorithms apply psychoacoustic filters that filter out signal components that our ears cannot perceive as the documents claim. Frequency and time masking is performed, i.e. information is deleted and cannot be recovered. MD use a fixed setup and their algorithm is proprietary (not open) while for MP3 the bit-rate can be chosen and its algorithm is documented openly.

In MP3 the bit-rate determines how good the original waveform is approximated. For speech 192 kbps is in general sufficient to achieve acceptable listening quality. The MD recorders re-synthesize the signal to make it externally available, i.e. computer captioning is done via an analogue outlet which also means a 3 dB signal quality decrease. For MP3 algorithms are available that turn the compressed representations into linear PCM representations. Analysis results (Campbell, van Son, Wittenburg) have shown that most of the usual linguistic analysis operations such as pitch extraction and spectral analysis can be done without getting large deviations compared to analyzing the original waveform. Nevertheless, due to the loss of information it is strongly recommended for archiving purposes to use the best recording quality possible, since we don't know what future generations may want to do with the material. It

has to be mentioned that in many cases the transformation of compressed formats to other compressed formats will not be without introducing severe artifacts.

Audio File Formats
There were a couple of file formats used to capture audio information such as NIST, AIFF etc. The de facto standard today is the **WAVE format** (.wav). It basically specifies how chunks of data can be read, in particular the format chunk informing about parameters such as the sample frequency etc and the data chunk containing the data in a specific byte order. WAVE is a particular sub-format of what was called RIFF (Resource Interchange File Format) which was created for different sort of applications. In fact all major programs support the WAVE format and there are converters to other formats such as AIFF etc. In practically all cases WAVE formatted files contain linear PCM data (differing resolution and sample frequency). Therefore, it is used as a synonym for linear PCM which is not fully correct.

Image Formats
For still images the difference between encoding and file formats is not obvious for the most cases, i.e. an encoding standard often also covers a file format. Therefore, we will not make this distinction here.

For still images also a number of encoding schemes are well-known. **TIFF** (Tagged Image File Format) is not standardized, it is more of a framework where different sub-communities have created their TIFF standard. Each manufacturer of high-resolution scanners produces his own TIFF version. Although it allows storing of compressed images as well, it is in general used for encodings that are comparable to PCM for waveforms. A picture is optically mapped to a lateral sensor that has a certain spatial resolution (number of image points in x and y dimensions yield a matrix of pixels) and for every pixel color and brightness are represented by a number of bits[4]. Therefore TIFF in general stands for uncompressed representations of image information. The major programs can handle a variety of TIFF formats although for specific versions (for example LANDSAT images) special viewers and converters are necessary.

Most popular is the JPEG format *(Joint Photographic Expert Group)*. It stands for a certain way to compress image information and for a file format. It makes use of discrete cosine transformations and the compression is achieved by cutting off high frequency components. Therefore sharp edges (lines) are smeared out. The compression factor can be chosen and of course the compression is lossy. Since all still cameras and most programs support JPEG this format is the de facto standard. Therefore, archives have to accept it. Since the format is openly described conversions to other formats are easily possible.
There are a few other but less important formats such as **GIF** (Graphics Interchange Format) and **PNG** (Portable Network

---

[2] "DAT Recording" is often used as synonym for good quality linear PCM recording. This is completely misleading, since the term "DAT" refers to a tape format (Digital Audio Tape). Correct is that for example the Sony DAT recorders used electronic circuits that generated high quality linear PCM (16 bit, 44.1/48 kHz). New types of recorders such as Flash-Card Recorders also support this type of high quality recording settings.

[3] Here the perception capacity of the human ear is used as indication instead of using human production characteristics.

---

[4] There are different color and brightness encoding schemes such as Grayscale, Pseudocolor, RGB, YcbCr and CMYK. This paper cannot go into the details.

Graphics). GIF was very popular at the start of the Web, since it is a highly reduced format. It just has 256 color encoding levels, does lossless compression and can be transferred very quickly via Internet. Since the owning company wanted to get money for every web-graphic, since the representation of color is so limited and since the network speeds increased it lost its importance. PNG supports lossless compression, supports a large color depth as JPEG and has a number of other excellent features making it very attractive. However, it is not very well supported by hardware and software builders.

JPEG2000 is a new standard that is intended to overcome some of the drawbacks of JPEG. Its compression is based on modern wavelet technology and therefore more optimal than for JPEG. It is fully specified in the mean time and aside to the core definition it covers various extensions such as for motions, file formats, APIs etc. Yet there is not much software that supports JPEG2000.

We should mention the **SVG graphics format** which is used to represent scalable vector graphics and is supported by W3C as a web-standard. ELAN makes use of this format to identify graphical shapes.

Video Encoding Standards
Uncompressed video which would be comparable to linear PCM in the audio world would amount to more then 250 Mbps, i.e. 30 minutes of video would require about 100 GB of storage. These two numbers indicate that uncompressed video still means a too heavy load on our current computer machinery.

Most relevant are the codecs worked out by the **MPEG group** (Moving Pictures Expert Group). They all do compression in the spatial and time domain and allow to define the amount of signal reduction. They also combine image and audio waveform encoding and define the packaging of the information stream such that decoders know how to unpack the stream and re-synthesize perceivable information. **MPEG1** was the first compression algorithm of this type and allows bit rates between 1 and 3 Mbps (mega-bits-per-second). In the spatial domain the discrete cosine transformation is applied to compress the signal. In the time domain group of pictures are defined that include a keyframe (full pixel representation) and a number of frames (prediction- and bi-directional-frames) that are highly compressed over time. They encode difference values between frames in a tricky way. PAL video[5] is delivered with two interlaced fields both covering half of the image and fields are sampled at a time resolution of 20 msec. MPEG1 only digitizes one of the frames, therefore it offers only a pixel resolution of 352*288 (also called SIF) and a time resolution of 40 msec. The resulting bit stream can be handled by CDROM drives which is the reason why MPEG1 was used for CDROM technology.

However, due to MPEG1's heavy reduction, its relatively poor quality and its tricky encoding scheme over time **MPEG2** was defined. It is based on similar compression principles, but

---

[5] NTSC Video in MPEG1: 352*240

encodes every field. Therefore, the original spatial resolution for PAL video is 704*576[6] and for time 20 msec[7]. It is widely used by the media industry for editing machines and was accepted as a kind of backend format[8]. Therefore, it is interesting for archiving purposes. MP3 is used for audio encoding in MPEG2.

MPEG4 is the last invention of the MPEG group[9] and is in so far very different from the first two ones, that it is more of a framework for decoding and merging several different streams of media information and supporting user interaction. It comes with an improved compression algorithm for video encoding and is designed for web-based applications. Although there already codecs around there is not so much software yet supporting MPEG4. It is mainly used for web-streaming[10] purposes. It allows to set bit-rates from 500 kbps on.

Many programs support the MPEG codec lines. However, writing software for proper decoding still seems to be a difficult issue, since it still occurs that programs or new program versions don't handle MPEG streams accurately. Video encoding requires to set a large number of parameters. The MPI team provides a template with default parameters that can be used with Adobe Premiere for example. There is a trend in TV and media industry to use MPEG2/4 with only I-frames, i.e. no intermediate frames with compression over time.

There are a few other codecs that are supported by some programs. **Cinepack** was used in old Quicktime versions, however, cannot be recommended anymore due to its bad quality. **Sorensen** is another codec, but not used very frequently. There are also proprietary codecs as provided by **RealVideo** and **Microsoft**. **DV** (Digital Video) is another very popular digital format. It was created from Sony for their digital cameras, but it is proprietary. These codecs don't play any role for archiving purposes such as in DOBES. The following table indicates again the data rates for the most important codecs (typical values, all for PAL).

	uncomp video	DV	MPEG1	MPEG2	MPEG4
bit rate in Mbit/sec	> 250 Mbps	35	1.5	6	0.5 - …
1 hour recording in GB	>100	16	0.7	3	0.2 - …

Video File Formats
The encoded video stream has to be packed into a file format. Here we only can briefly mention some of them without describing

---

[6] NTSC Video in MPEG2: 704*480

[7] There are various other resolutions supported which will not be reported here (HDI, QSIF, …).

[8] MPEG2 4:2:2P is the standard that was agreed by the big media industry for their "mainstream production lines". The numbers 4:2:2 point to the relation for encoding brightness and color information (Y,U,V model). MPEG2 is the standard for DVD technology.

[9] MPEG7 and MPEG21 are not about codecs and file formats.

[10] In contrast to the download option where data is first downloaded and stored on the local machine, streaming means that data is directly presented on the screen while being sent.

the details. MPEG (.mpg) comes along with its own file format. Another very popular format is AVI (.avi). While .mpg format only includes streams encoded with the MPEG codec, AVI can include all sorts of codecs, i.e. saying that something is an AVI file does not say anything about quality and other characteristics. DV streams for example can be embedded in AVI.

APIs for Complex Streams
There is a group of formats that go beyond a simple file format such as **Quicktime** and **SMIL**. QT and SMIL (which is a recognized standard by W3C) both can contain several tracks of different information types that are linked by cross-references. This would for example allow to store video and its synchronized subtitles. An API (Application Programmer Interface) is mostly provided that tells the user how the different objects can be dealt with. Media can be encoded in different ways, i.e. saying that a video is in QT format does not say which codec is used. QT is very popular since it was pushed by Apple and since there are QT Players for Windows and MAC OS. QT is supporting a number of codecs. SMIL is very similar, an open standard and will be supported by the major web-browsers. MPI will probably offer subtitled video presentations via the web by providing SMIL documents.

MXF (Material Exchange Format) will become a very important format for media exchange, since the big TV and media companies agreed on it. It includes data and metadata and can codify different tracks of related streams. It will play an important role for future archiving, i.e. we expect that the DOBES archive at a certain moment may have to turn over to MXF.

Summary
This brief overview may have given a better insight to what currently is used and recommended. It is impossible to write a comprehensive document about these issues that will not cross the 100 pages boundary. Therefore, it is always recommended to speak with experts, since much knowledge especially about tools resides in their heads only. We can't speak about a stable situation in the area of codecs and formats, since within a 5 years time new suggestions for codecs and formats will be made. Whether such suggestions or those that are already around will become standards relevant for the documentation work and for archivists cannot be predicted. There is a clear trend to uncompressed representations, but technology has to support this.

Send contributions for the next
issue to: **LAN@mpi.nl**
before **March 31, 2004**