

# Annotations, Formats and Data Types in the DOBES Project

Peter Wittenburg, Hennie Brugman, Daan Broeder  
Max-Planck-Institute for Psycholinguistics  
peter.wittenburg@mpi.nl

Paper presented at the workshop on  
Web-Based Language Documentation and Description  
12-15 December 2000, Philadelphia, USA

## 1. Introduction

The Max-Planck-Institute is a center where scientists continuously create multimedia/multimodal language resources based on field and laboratory recordings. In contrast to a few years ago the resources stored on computers now have multimedia extensions, i.e. the audio and video signals<sup>1</sup> are immediately accessible by the researcher as well. The challenge of combining media signals and annotations has made a big deficit very apparent: we lacked a clear organizational scheme and infrastructure to create, store and access all those resources that are of interest. It was understood that the individual researcher was not capable of carrying out this task in a way that could guarantee re-usage even within the institute. The corpus infrastructure generated by about 30 researchers was collapsing into chaos.

Therefore, about two years ago the MPI decided to build up a well-organized and distributed infrastructure operating both in the Intranet and the Internet. This infrastructure should offer environments that allow us to integrate the resources from the moment of its creation, while hiding physical structure (disks, directories) from the user, and allowing the user to operate in a conceptual domain. This domain is defined by terms such as “languages studied”, “gender and age of informants”, “genres and tasks being chosen for the recording” etc. The user should be able to access resources and operate on them independent of their location given that he has the rights to do so. Therefore, the institute started to design an environment that is mainly identified by the following components: (1) A workflow scheme which describes the individual steps starting from the planning phase, (2) a redundant infrastructure for digitization, (3) a repository of meta descriptions which allows the user to easily find the right resources and to execute operations on them [1], (4) a format-independent and Internet-capable multimedia tool [2], and (5) a reliable archive infrastructure.

Recently the MPI got the task to house the DOBES project<sup>2</sup> (Documentation of Endangered Languages) [3] to introduce and further develop all the components just mentioned. DOBES from the beginning was intended to be a web-based multimedia archive.

In this paper we will explain the main components applied in setting up the MPI and especially the DOBES archives, with discussion of what can be used to set up a more general framework for similar archives. Further the paper will present the data types and formats the linguistic projects produce and how they are integrated into the archive. Also the resulting formats and integration suggestions can be seen as contribution to a general open framework.

## 2. Workflow Scheme

The amount of resources at the MPI require a professional solution in so far that a support team takes over the setup of the archive, i.e. take the raw material, process and organize it, and deliver the ready-to-go results to the scientists. This split in labor required a formal workflow scheme for each fieldwork project. Such a scheme was discussed in the MPI and now also serves as basis for our activities within the DOBES project.

The workflow scheme shown in figure 1 is the common basis and is adapted to the specific needs of the individual projects. It basically says that a set of tapes is produced where each tape can cover several recording sessions. Per recording session the researcher is required to produce a metadata description, otherwise the raw material will not be accepted as it cannot be

---

<sup>1</sup> The extensions are not limited to audio and video signals. It could be eye tracking or brain scanner data as well.

<sup>2</sup> The DOBES project is being funded by the Volkswagen Stiftung.

integrated into the archive. The archive administrators then get the set of tapes with standard labels and a set of associated meta descriptions. The meta descriptions for the DOBES project are created following a set of metadata elements, which is in accordance with the IMDI metadata proposal [4,5]. The meta descriptions include the tape label, the session number on the tape, the start and the end time of the session, which are used to automatically control the digitization process.

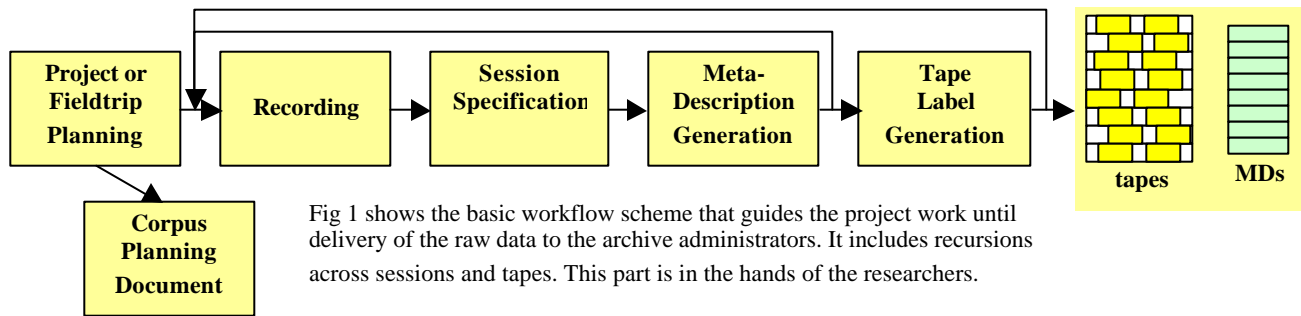


Fig 1 shows the basic workflow scheme that guides the project work until delivery of the raw data to the archive administrators. It includes recursions across sessions and tapes. This part is in the hands of the researchers.

Dependent on the circumstances of individual projects also adapted workflow schemes are used where for example the raw data on tapes is first sent to the MPI for digitization. The digitized material is sent back per Internet or CD/DVDROM. The main advantage of this scheme is that there is no time code problem anymore. The segment boundaries extracted with the help of software can be used to generate the session files by simple scripts.

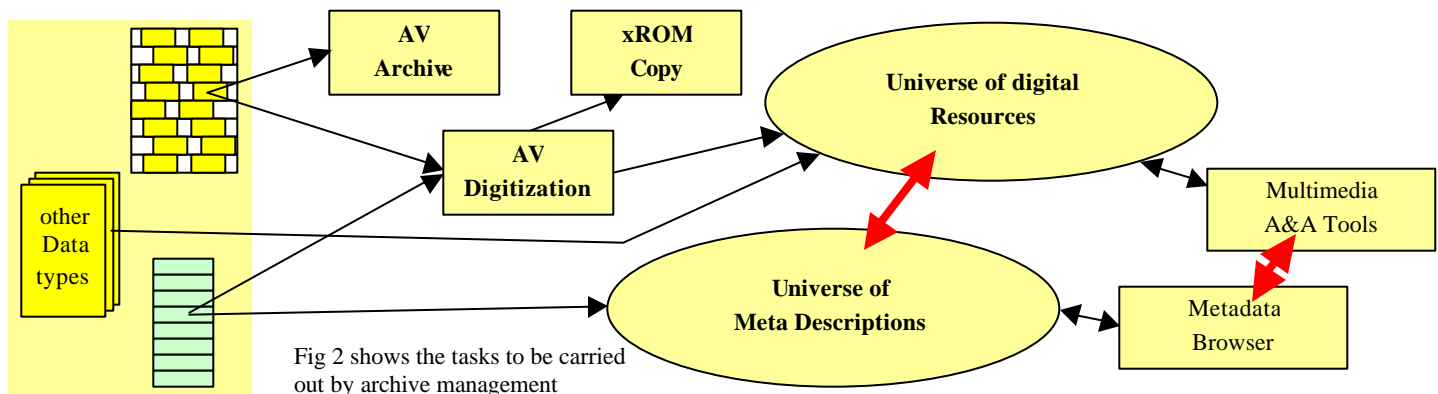


Fig 2 shows the tasks to be carried out by archive management

The workflow scheme also makes statements about the further processing steps that are mainly done by the archive managers. During this process sessions are segmented and the deletion of irrelevant parts of the recordings is therefore automatically performed. The original tapes will be stored in general in the AV archive. The digitized material and corresponding data types such as transcriptions, annotations, field notes etc are all integrated into a structured “universe of resources”. The meta descriptions are integrated into the linked hierarchies, which make up an easily browsable & searchable metadata “universe”. When looking for some resources the user is mainly confronted with this metadata universe. Having found the right resources the user can directly start operations such as searching for patterns in a set of resources or starting a multimedia annotation tool from the EUDICO tool set. Both universes can be distributed across the Internet. Also the major tools involved (Browser [6] and EUDICO [2]) are prepared to operate in such distributed worlds.

### 3. Infrastructure for Digitization

One of the keystones of our decisions was to create an all-digital archive to make all data including the media data immediately available for everyone via the networks. This is the only way to deal with media data in a feasible and reliable way in the future and to make access convenient to the user. The key elements of a digitization infrastructure are efficient digitization facilities and proven standards. For all media which we are supporting (audio cassettes and tapes, DAT audio, Hi8, DV, ...) we give guidelines to our users how they have to do the recordings such that the material can be digitized without side effects. Based on these guidelines schemes were defined about how to digitize audio and video tapes as efficiently as possible. Mostly off-the-shelf equipment and software is used for digitization. However, to efficiently organize this process a number of Perl scripts have been added.

The results are media files which are handled again in accordance with the workflow scheme and the meta descriptions, which we have got. The following conventions have been defined as standards within DOBES and for the MPI:

Audio Files: wav formatted files with sampling rate either being 44.1 kHz or 48 kHz

Video Files: MPEG2 (3-6 Mbps variable bit rate) encoded files will be generated to assure that a high quality signal is stored<sup>3</sup>.  
 MPEG1 (1.5 –3 Mbps variable bitrate) is generated in parallel as an temporary format in house and as a format that can easily be used via the networks.  
 In parallel also wav files are being generated especially to be able to show speech waves and allow the user to calculate parameters such as pitch contours.

## 4. Meta Description Universe

As should be clear from the workflow scheme the meta descriptions are essential to organize the resource world such that the user can easily find the data he is interested in. In contrast to the resources themselves the meta descriptions should be available to all researchers to inform each other about what is available. Meta descriptions are integrated into hierarchies that are meaningful to the users.

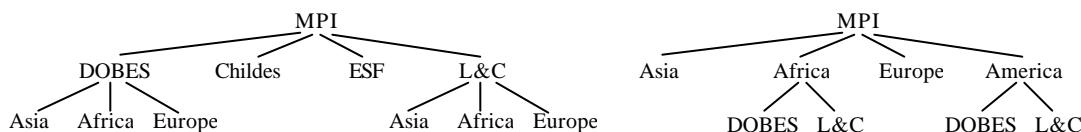
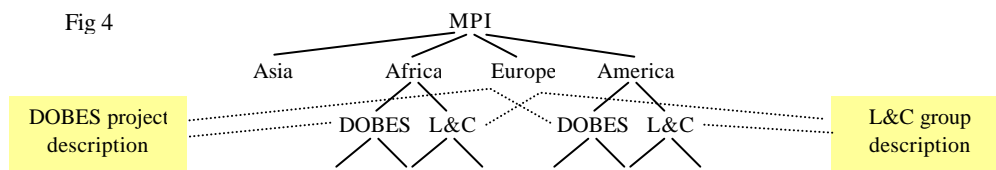


Fig 3 shows two different type of meta description hierarchies.

Several different parallel hierarchies can be used to guide the user when browsing through the linked descriptions. Each user will have own preferences and the preferred structure also depends on the task that has to be accomplished. It is possible to allow a user to create own hierarchies.

Each node in the hierarchy itself represents a meta description. It can incorporate useful information itself or it might be the place to which explanations, fieldnotes or photos might be attached. In fig 4 it is shown how a description of the DOBES project or of the Language&Cognition group at MPI could be attached. When browsing in the metadata universe the user might want to see this information and the browser should render it. The format of such descriptions is either ASCII or HTML at this moment. Later XML coded files could also be acceptable. The same is true for photos or images which have to be in either gif or jpg format. The description files could, of course, contain links themselves that open many possibilities to add rich information.



At the bottom of the tree there is information about the individual sessions. Here one can find the elaborate meta descriptions, which describe a session, its formal characteristics, its content, its participants, its annotations and others. They also contain the references to the real resources, such that the user can directly activate tools to operate on them, i.e. create new annotations, compare video segments, perform a query, or simply to copy it to a CDROM. In doing so the physical location of the resources is completely hidden to the researcher.

All meta descriptions are contained in structured XML files. The DTDs of the session MD files and that of higher level nodes in the hierarchies are attached in appendix 1. Within the DOBES project a metadata set is used which is a subset of the set that has recently been proposed by IMDI [5]. When there is a widely agreed standard MPI will provide scripts to transform the existing meta descriptions.

To create the meta descriptions MPI has developed a specific user-friendly editor, which is available as a Java program or in future as a down-loadable applet. Also the browser is written in Java, available as program or applet, and can operate with meta descriptions, which are residing on different sites in the Internet. MPI currently adapts both tools to the DOBES core set. IMDI is expecting a first draft of the IMDI metadata set in the spring 2001. MPI will then immediately adapt its editor and browser to the agreements and make both available for all interested people. The editor directly creates XML files, but also the DTDs (or XML schema definitions) will be openly available. For a starting phase the MPI would also be willing

<sup>3</sup> It was decided that for technological reasons DV or MPEG2 with I-frames only etc cannot be used, since the data rates are simply too high for current technology (factor 10).

to house meta descriptions created by other institutions, but here definitive solutions have to be found within the community.

## 5. Multimedia Software Toolset

Another key component in the web-based documentation and description of languages and for other similar tasks within the MPI and the DOBES project is the availability of a multimedia annotation and exploitation tool which operates in a distributed environment also. The MPI is currently working on EUDICO. It is expected that a first version covering all relevant components will be ready in summer 2001. Currently, a version is running at test sites, which has most of the core functionality with the exception of distributed and protected data entry. This data entry component is designed such that researcher X sitting at a location x could do collaborative work with researcher Y sitting at location y. It could mean that while X is creating the transcription Y could work on gesture, both looking at the same video fragment which is sent across the network and both immediately seeing what the other is coding. They even could exchange comments that the colleague can visualize. Tests with sending video fragments via the Dutch scientific network indicate that this sort of collaborative work is already possible in many countries. However, EUDICO can also run on current state-of-the-art notebooks as a local version – possibly reading the video information from an attached DVDROM.

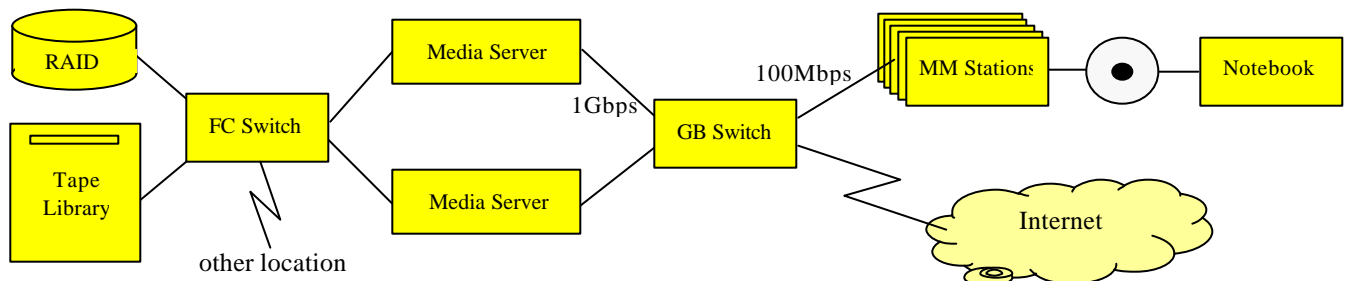
Besides its Internet capability EUDICO has the following major features: platform independence due to 100% Java, corpus format independence due to a kernel which incorporates an Abstract Corpus Model, several synchronized text, sound and video viewers, viewers which show the time alignment between different independent annotations, viewers which show dynamic subtitles, a generic and powerful search tool operating on all connected corpora, a media streaming component which sends only the required video frames via the network, and import/export modules for SHOEBOX, CHAT, Transcriber, relational databases, and specific XML-based corpora, and distributed as well as local operation even on notebooks. The MPI is currently working on the following functionality: integration of a type definition tool to define the type and structure of new annotations, integration of UNICODE to allow users to enter and visualize non-ISO-latin characters, an improved version of the search tool, integration of a professional sound analysis tool, integration of a multimedia lexicon component, and integration of a general XML-based export format which could serve as general and open interchange format.

EUDICO can be configured as a coherent tool set for the typical tasks which fieldwork researchers have to carry out. It will have all basic functionality. But it will also offer interfaces to hook up other tools or to let people for example define their own import component to adapt to another corpus format. EUDICO will be tightly integrated with the metadata environment. EUDICO will be further developed not only for the MPI, but also under contract for a number of external projects.

MPI is willing to offer EUDICO as well as its other tools for linguistic applications [7] to the academic community. The formats supported by EUDICO are well-known and described [8,9,10,11]. It is intended to define a general and open exchange format in tight synchronization with the Talkbank project. It is expected that this will become a powerful and widely accepted format in the language resource community.

## 6. Multimedia Archive Infrastructure

Another key component of Internet-based language documentation is a multimedia archive infrastructure. Storing multimedia information requires professional storage management, powerful media servers, a powerful switching network, and a high capacity connection to the backbone of the national scientific network. The storage architecture is determined by the need to handle large data volumes (at MPI currently > 400 GB), since modern language archives incorporate digitized audio and increasingly video data (1 h MPEG1 video = 1 GB data). Further, archives need reliable data storage especially when speaking about archiving material about endangered languages for many years. Reliable data storage is based on



automatically maintaining two copies at different locations and moving to new media generations. To provide the necessary media streams in a mode free of jitter to the Intranet- and Internet-based desktop machines a scalable media server configuration and fast network components have to be setup.

At the MPI such an infrastructure has been established during the last years to house the multimedia language resources of the institute and in future of the DOBES project. The MPI is willing to share this data with others, as long as the access rights are not violated and the costs can be limited. Setting up an infrastructure, which can provide media streams to the user desktops, requires the availability of adequate software. EUDICO was designed such that it supports sending media fragments on demand via the networks. The MPI is willing to share the knowledge to others of how to set up such an archive structure.

## 7. Data Types, Annotations, and Formats

Far most essential for a distributed Internet-based archive infrastructure is the easy exchangeability and re-usability of the resources, be they media data, annotations, lexicons, or other textual material. These requirements have been cited very often, however, technology develops very fast, new standards are continuously created, and users are increasingly given more possibilities to satisfy their needs. Nevertheless, to prevent the emergence of a chaos we have to keep taking efforts to create adequate guidelines and standards. But there is no doubt that we will be faced with a number of different formats, tag sets, and encoding schemes in the future too. To achieve interoperability we will have to take also efforts to develop mapping schemes at different levels.

Within the DOBES project the data type and format issues have been addressed (for an overview see appendix 2). The set of data types is not fixed, there will be annotations of multimedia data, lexicons, wordlists, field notes of various sort, grammar notes, notes of the sound system of a language and others. With respect to the annotations we hope that the Atlas Interchange Format will become so powerful that all relevant information can be coded. First attempts to define a general lexicon format are on the way. With respect to the other data types we are faced with a wide variety of encoding types and formats. Users prefer to write such notes for example in MSWord and save them as DOC files. It is intended to convert all these data to either plain text or html descriptions and to make them available by including links in the meta descriptions. The metadata browser has the capability to render such information. From our experience it is too early to try to define standards unifying the syntax and the semantics of these other data types.

Also the list of annotation tiers describing the linguistic phenomena to be found in the media data is open. In DOBES some tiers are mandatory (orthographic, phonetic/phonemic, morpho-syntactic, English glossing, English translation), but it was clear that there will also be Chinese or French translations, annotations of tones, and others more. Even more different annotation layers are used at the MPI. Important for archiving is a certain degree of interoperability (searching for certain morphological phenomena). Not only the same tag set but also the same encoding should be used to describe these linguistic phenomena. Due to the differences between the languages there will not be a complete overlap, but it was agreed for example in the DOBES project to unify as much as possible for example for the morpho-syntactic and glossing tiers. The procedure followed is that each team (per language) describes the comparatively small set of encoding elements they are using. The MPI team will try to do a mapping between all and provide lists that are also in accordance with for example the EUROTYPE suggestions. The set of elements used is open.

The area of data types and formats is still evolving and often linguistic theories or personal impressions are involved. We don't understand the ontology yet fully and we have to treat the material as such, i.e. make it easily available and render it as text partly structured by the user himself in an idiosyncratic way. Some examples of typical data types and its formats will be visible via the DOBES web-site. Also the encoding elements used for the morpho-syntax and glossing tiers will be made available.

### References

- [1] [www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)
- [2] [www.mpi.nl/world/tg/lapp/eudico/eudico.html](http://www.mpi.nl/world/tg/lapp/eudico/eudico.html)
- [3] [www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)
- [4] D. Broeder, P. Suihkonen, P. Wittenburg: "Developing a standard for meta-descriptions of multimedia language resources. Proceedings of this workshop
- [5] [www.mpi.nl/ISLE/documents/docs\\_frame.html](http://www.mpi.nl/ISLE/documents/docs_frame.html)
- [6] [www.mpi.nl/world/tg/lapp/browscorp/browscorp.html](http://www.mpi.nl/world/tg/lapp/browscorp/browscorp.html)
- [7] [www.mpi.nl/world/tg/lapp/lapp.html](http://www.mpi.nl/world/tg/lapp/lapp.html)
- [8] [childes.psy.cmu.edu/](http://childes.psy.cmu.edu/)
- [9] [www.sil.org/computing/catalog/shoebox.html](http://www.sil.org/computing/catalog/shoebox.html)
- [10] [www ldc.upenn.edu/mirror/Transcriber/](http://www ldc.upenn.edu/mirror/Transcriber/)
- [11] [www.mpi.nl/world/tg/lapp/lapp.html](http://www.mpi.nl/world/tg/lapp/lapp.html)

Appendix 1: DTD of Meta-Description files

Appendix 2: DOBES overview about Data Types, Annotation Tiers, and Formats