elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 1*. Editor: Peter K. Austin

The DOBES model of language documentation

PETER WITTENBURG

Cite this article: Peter Wittenburg (2003). The DOBES model of language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, vol 1*. London: SOAS. pp. 122-139

Link to this article: http://www.elpublishing.org/PID/011

This electronic version first published: July 2014

The DOBES Model of Language Documentation

Peter Wittenburg

The paper presents the agreements made in the DOBES programme,[1] the state of the documentation and archiving work and discusses a number of technological problems that the programme is faced with. This paper is not intended to give a broad account of the linguistic aspects of the work, but focuses on technological aspects and those that have to do with the archiving task.

1. Scenario

The major goal of the DOBES programme founded in 2000 is the documentation of endangered languages, i.e. languages that will become extinct within a few decades. Documentation means to record and describe languages so that later generations can reconstruct them. Such a documentation contributes to preserving an essential part of human heritage. The scenario that is given for the DOBES programme is depicted in figure 1.

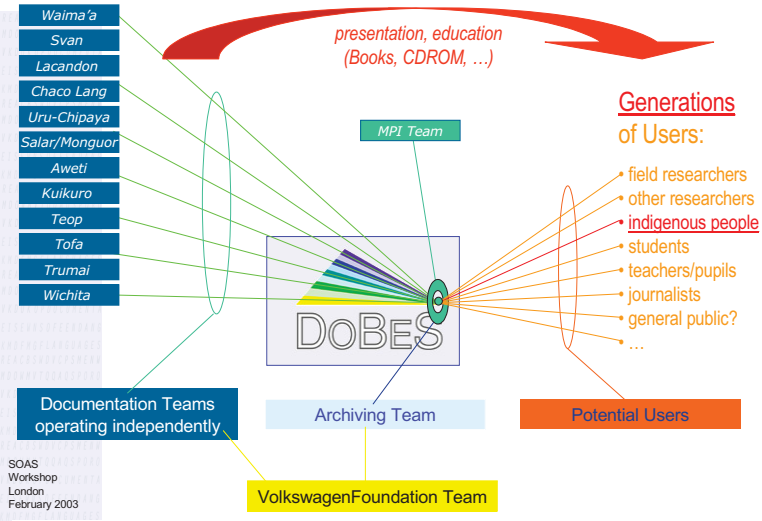


Figure 1 describes the essential construction of the DOBES program. It consists of the documentation teams that produce material and send it to the archiving team. The archiving team has to offer the data to interested users and preserve it. The team of the Volkswagen Foundation helps to run the whole programme at an administrational and organizational level.

[1] The DOBES programme is entirely funded by the Volkswagen Foundation.

Currently, the DOBES programme includes 21 teams consisting of linguists, ethnologists, musicologists, ethnobiologists and others from related disciplines, who look at about 30 languages[2]. The languages selected within the DOBES programme show a worldwide distribution. Each team works independently since the specific characteristics of the language in focus and the specific circumstances of the fieldwork at each field site are rather unique. However, to realize the idea of a central archive requires a certain amount of agreement about the methods to be used and about the formats to work with.

One of the basic concepts of the DOBES programme is that the documentation teams will make copies of their data available to the DOBES Archive that is housed at the Max-Planck-Institute for Psycholinguistics. It should be emphasized, however, that DOBES encourages the teams to exploit their data in a way that may serve different needs, in particular the needs of the language community. Only the teams know which types of material the communities will need in order to teach young children, to attract community members to take active part in the documentation or to create CDROMs to interest governments and others for their work, and to produce results from linguistic analysis etc.

The archive has to store the data in a standardized form and to make it available to all possible interested groups such as researchers, journalists, teachers, students and others. The archive has to offer its content to users who may want to work with the material, for example, to add new annotations or to create comparative studies between different languages. Another main task of the archive is to look for strategies that will increase the probability that future generations will still be able to access the data. This is one of the big challenges for Information Technology (IT) these days, since our storage media last only for a very short time compared, for example, with the clay tablets that were used by the Sumerians.

Finally, it should be mentioned that the DOBES programme is completely funded by the Volkswagen Foundation (VWS). However, the VWS restricts itself to a passive role. It will not interact with the documentation and archiving decisions. The VWS is establishing boards that include well-known scientists from the field who will decide on new projects and observe the state of the work. It is the Steering Committee that is being elected from members of the teams that will take initiative in case of problems of general concern. Therefore, we can conclude that DOBES is a programme, which is organized rather bottom up, but in which the central archive requires a number of agreements.

2. Agreements

As mentioned already a number of agreements between the teams and the archivist where necessary for developing an approach as coherent as possible. They fall into two different

[2] In the meantime additional teams have been accepted into the DOBES programme. They work on the following languages not yet mentioned in the figure: Mave/Katxuyana/Bakairi, Hocank, Tsafiki, Chol, !Xoo, Akhoe-Hai//om, Iwaidja, Chintang/Puma, Marquesan

categories: (1) agreements about documentation aspects and (2) agreements about technical aspects.

2.1 Linguistic Agreements

It is one of the basic decisions in the DOBES programme that languages cannot be documented without referring to their cultural background. Language is an essential part of culture, since it is the language that allows people to communicate about economic, social, cultural and other aspects relevant for the survival of the community. For example, if a speaker talks about the many names given to a certain animal it would be excellent to explain why this is so and this should be supported by video recordings. Another example could be the description of how people build houses in traditional ways. It would be excellent, if the description could be associated with a video showing how the building process is actually implemented.

Further, it would be desirable if the documentation could show how for example the sounds of the language are realized (behavior of articulators) and how gestures are used to support verbal interaction. Therefore, the DOBES program stresses the need for multimedia recordings (audio and video) as the basis of any documentation effort.

In addition, the DOBES project stresses the multidisciplinary approach to language documentation. It encourages the teams to not only include linguists who work on the proper linguistic analysis of the language material, but also to include for example ethnologists to investigate relevant aspects of the culture and musicologists to document and analyze the essential part of heritage that is manifested in songs, dancing etc.

In several workshops during the pilot phase the teams discussed a number of issues that should guide the documentation work. Rough guidelines were established for the selection of text types and genres. For the annotation of the recordings it was agreed that there should be two tiers for all material: (1) a tier containing either an orthographic transcription or a phonetic transcription (in the case that no standard orthography is available) and (2) a tier containing a translation into a major language. Further, it was recommended to create translations into a local lingua franca and into English if the “major language” is not English. To keep the documentation task tractable it was agreed that only a small part of the recorded material can be subjected to further in-depth linguistic analysis. It was left to the teams to decide which type of tiers such an in-depth linguistic analysis should include. A comprehensive annotation approach was proposed with the Advanced Glossing model [1].

Moreover, it was agreed that all terms and conventions used during glossing and analysis (such as the morphosyntactic terms used) should be documented carefully. With respect to lexicographic work a few basic principles were agreed upon, such as that the work should be topic oriented. In addition, some classical linguistic material should be

provided such as sketch grammars, field notes, notes about the sound system of the language, etc. A conference will be held in 2004 to summarize the guidelines and to collect the experiences made by the teams.

2.2 Technical Agreements

A number of agreements were worked out that describe the standards relevant for the archive and for the interaction between the documentation teams and the archive.

It was agreed that the archive should be based on

- XML for the structuring of textual data[3],
- PDF or HTML as alternative formats for documents wherever useful,
- UNICODE for the character encoding,
- MPEG2 as backend format for the video recordings[4],
- MPEG1 and MPEG4 as frontend formats to be delivered for the analysis,
- linear PCM 44/48 kHz encoding of sound signals[5] (WAV file format)
- JPEG and TIFF as encoding formats for images[6]''

A broad and open discussion took place about tools that could be used within DOBES for documentation purposes. Based on the input from many teams a set of tools was recommended to minimize the conversion task. This list of tools is dynamic and it will be subject to further discussions, since new tools may enter the scene that turn out to be very useful and that don't create complicated conversion problems. The list covers programs such as SHOEBOX [2], TRANSCRIBER [3], PRAAT [4] and ELAN [5]. As any other documentation project too, DOBES was confronted with existing data and with established routines of the researchers. As a consequence the archiving team was confronted with other formats such as WORD DOC, WORDPERFECT and IPA [6] annotations in PRAAT. For all these formats conversion programs and scripts had to be created that also had to take care of character conversion. Since several researchers were using WORD as a tool to

[3] Since EAF [13] is seen as a flexible XML-based annotation format, the archive will use this format for storing annotations, i.e. all other annotation formats will be converted to EAF.

[4] DV (Digital Video) is a very popular video format used by many camcorders. However, it has a high data rate and it is a proprietary format.

[5] A considerable amount of time was devoted to the debate whether MP3 or ATRAC (Minidisc) encoded speech should be accepted. It was agreed that high quality speech encoding should be applied whenever there are no special reasons (e.g., due to the field conditions) that make it necessary to use for example Minidisc devices.

[6] With respect to the encoding of images it was understood that neither JPEG (as a compressed format) nor TIFF (as not being fully standardized) are ideal formats.

create lexicon entries and annotations, special attention was given to the WORD-to-XML converter. It includes a small language that allows the researchers to describe the structure of their WORD-document and to associate XML-tags with such structural elements.

Since SHOEBOX plays an important role in all documentation work, it was important to create converters and import/export methods for SHOEBOX. Therefore, ELAN (a tool developed by the MPI archive) can import and export SHOEBOX format. There is also a converter between SHOEBOX and TRANSCRIBER, since the latter is a very popular and efficient tool for transcribing sound files.

To organize the archive the IMDI (ISLE Metadata Initiative) metadata approach was chosen. In several discussions the specific requirements from field linguists were discussed and considered during the general discussions about the IMDI metadata set [7]. As one result, it was agreed that every team will create a canonical tree that will determine the structure of the corpus as proposed by the researchers[7]. This structure is also used in the communication between the teams and the archivist to place recordings and other types of resources provided by the teams.

Most important for the success of the project were in-depth discussions about the workflow, i.e. the steps and procedures to be taken during the interaction between the teams and the archivist. In the pilot phase several discussions took place with almost all teams to achieve a better mutual understanding. On the one hand the archivist had to understand the actual work situation of each individual team and on the other hand the researchers used to working individually had to better understand the requirements resulting from the archiving task. For every team a detailed workflow document was generated collaboratively to describe the labels used and the procedures implemented. Training courses for the new teams were given that elaborated on this issue, since mutual understanding was seen as the key to a successful interaction.

3. State of the Archive

Due to good workflow agreements and the high level of mutual understanding the archive was able to make good progress. Two principal workflows were discussed as indicated in figure 2:

1. The centralized workflow (the upper path) specifies that the documentation teams send their recordings to the archivist where the material is captured. Some tests showed that the best approach is to capture a whole tape (resulting in a Digital Master File - DMF) and to send the resulting MPEG-files via DVD back to the teams. These can then use software to define their sessions (linguistic meaningful units of analysis), which are further analyzed. The teams then simply provide

[7] The canonical tree is the tree maintained by the archive and used for management. In IMDI every user can create additional tree structures that are most suitable to the current work.

metadata descriptions to the archivist that contain (for each session) the exact begin and end time. Via batch processes the archivist can then cut the sessions in the same way and place the session files at the right location into the corpus structure.

2. The decentralized workflow (the lower path) specifies that the teams do the video capturing and all the definition and cutting of sessions in the field. In this case the archivist could get CDROMs[8] with the session material and appropriate metadata descriptions.

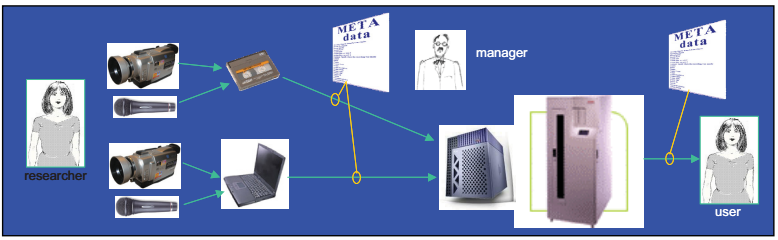


Figure 2 indicates the two major workflow principles that are used in DOBES. The upper path indicates the centralized and the lower path indicates the decentralized procedure in capturing audio and video and in interacting with the archive.

Both methods have their advantages and disadvantages. The centralized method was important in the pilot phase, since it turned out to be very robust and since the interaction with the teams was relatively simple. However, the cycle time is very slow, i.e. it takes time before the teams get the material back so that they can start working. The decentralized method has a short cycle time, which is especially necessary in situations where the teams collaborate with native speakers in the field to do the transcription and other annotations. However, it is more error-prone in so far as the capturing and conversion process from DV to MPEG has to be done very carefully, using the right parameter settings. Computer-based video processing is not yet as stable and trivial as sound processing. In future the decentralized model will be chosen more often. However, this requires powerful notebooks, careful data- and power management in the field and users who know what they are doing. Otherwise lots of problems can be expected to arise that the archivist will then have to solve.

By the end of 2002 the archive had grown to 350 Digital Master Files (i.e. about 350 hours of sound and video recordings), 561 sessions that were derived from these DMFs, 49 annotations, many photos and field notes, and some first grammars and lexicons. Of course, the first teams started in 2000 and will deliver their documentation results by

[8] One of the problems could be that CD-ROMs are too small. But currently no Notebooks can write DVDROMS. The teams could also send the tapes under certain circumstances.

the end of 2003 or by the beginning of 2004. Thus, it can be expected that annotations and other derived linguistic material will arrive in that time period.

4. Archiving Issues

As already mentioned the archivist has to deal with the task of how to best preserve the digital material for future generations. At the beginning of the DOBES project the MPI gave the assurance that it would make the collected information available for about 10 years[9]. This may indicate the dramatic situation for long-term archiving of digital media. We can point to three different aspects that have to be tackled: (1) the problem of short media lifetime; (2) the risk of failures and (3) the interpretation problem.

The best we can do to store large amounts of digital information in a safe way is to store it on storage media that have lifetimes of less than 10 years[10]. This formulation seems to be absurd, especially if we compare this with the cuneiforms of the Sumerians for example which we can still read although they were created more than 2000 years ago or if we compare it with old paper documents using types of material different from those used today which also survived for very long periods and which can still be read. However, we have to compare this decrease in media lifetime to other attributes such as the time necessary to copy the content. In this respect our current "technology" is much better than the old ones. So if we want to achieve long-term storage of huge amounts of data (compared with the amount of data stored in early books or even more extremely on clay tablets), given the short lifetime, we have to take care that we will continuously copy data to new storage media (continuous migration) that may even offer increasingly shorter access and reading speed and to make this process as cheap as possible. So if we want to classify the usefulness of a technology for achieving long-term availability, we have to look at the combination of parameters such as amount of data, copy time and media lifetime. In this respect our current technologies are not at all as bad as it may look at the first glimpse.

Another factor is introduced when trying to achieve long-term availability. While the clay tablets, once created, could survive without further intervention if they were stored under normal circumstances, we have to solve an organizational problem for our current media. The copying process has to be organized and funded, i.e. financial effort is needed

[9] Recently, the MPI made serious steps to offer the availability of the data for 50 years, which is still an extremely short period considering the historical dimension. It is expected that within this period new solutions will be worked out using so-called Data-Grid technologies.

[10] Some speak about lifetimes of 30 years for CD-ROMs for example. Comparatively these are still short time periods. But more important is the fact that CD-ROMs do not allow handling dynamic resources and that the 30 years specification is a fictive one, since the technology to read them may not be available anymore except for rather specialized institutions.

that will make data survival dependent on the political and economic situation, which can neither be anticipated nor influenced.

This leads us to the aspect of failures in the procedure of copying. Currently, DOBES data is stored on 5 copies: 3 archive copies are created dynamically and are up to date; one archive copy is created by generating snap-shots at specific moments, and at least another copy is in the hands of the researchers. But with the exception of the copies created at the archivist site much manual work is still involved and is therefore not reliable. Manual work is error-prone and is not a good concept to survive an economical crisis. So we need automatic algorithms to create these copies.

Further, to make the survival independent of political and economical situations we have to distribute the data in a controlled way across the whole earth. With the exception of a big crisis that has worldwide consequences one can assume that at least one copy will survive, which will then spread out when the automatic copying and distribution procedures are operational again. Computer science has to develop and test out data-grid technologies as soon as possible on a large-scale basis. These will take care of deciding which partners worldwide can be seen as trusted data centers adhering to the same set of ethical and legal rules, of automatically copying and distributing (parts of) the data to other places and keeping track of the locations where the copies are stored. Data-grid technologies have to be supported by worldwide agreements to protect their operation. This requires national commitments for funding them.

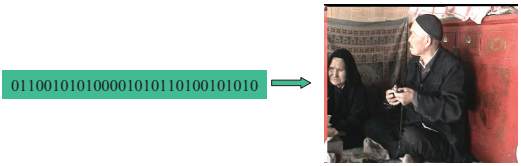


Figure 3 indicates the problem of how to interpret a stream of bits as stored on computers to produce for example a meaningful presentation.

The third aspect concerns whether future generations will be able to decode our encodings, i.e. to interpret our bit streams in the right way, as indicated in figure 3. If we assume that by continuous migration and worldwide distribution of the data we can guarantee long persistence times covering many generations, the remaining question is whether future generations will be able to interpret the bit stream stored so that, for example, an understandable movie can be generated.

Some experts argue that we don't have to care about this issue. This is one of two extreme positions. Specialists will find out how to decode the material, if there is a sufficiently large societal interest, even if there is no comprehensive documentation. In this way specialists were able to decode the information contained, for example, in the cuneiforms.

The other extreme position would be to assure that people at all times are able to interpret the bit streams with the tools they have available at that moment in time. This would mean, however, that every generation would have to solve the task of the migration of the encodings. Such a migration would include the transformation of file formats to new formats, the transformation of the encodings such as MPEG2 to something new, the adaptation of the documentation, since our descriptive language and in particular the procedures will change. Considerable amount of time and money would be involved and it is hard to see that societies at all moments in time will commit themselves to such tasks.

The interpretation and transformation costs would even increase if there were a great variety of original formats and encodings. Therefore, the DOBES archivist decided to build a coherent archive where all contributions are organized according to one clear schema (the IMDI metadata schema) and transformed according to a limited set of standards as described above. It is obvious that this policy also assures that at present the access to the material is simplified[11].

So in the DOBES programme we follow the "immediate way" of achieving coherence. This is in opposition to the "later way" where various formats and encodings are taken and where it is expected that coherence will be achieved at a later moment. Both approaches have their advantages and disadvantages. The immediate way requires a good synchronization with the documentation teams and they have to adapt their habits. The training effort is considerable, since the researchers who want to focus on the documentation work itself also have to learn how to use tools that have a potential to create formats that can be handled easily. Nevertheless, some transformations have to take place, i.e. some time and money has to be spent immediately.

For archives following the "later way" there is much less effort necessary to convince and train people, since they simply can take every type of digital material that is offered. Even the categorization can be done later. This implies that there is no coherent archive. It would cost time then for a user to discover suitable material. Since different types of formats and encodings will be found in the archive, the archivist and users have to learn how to use several tools when access to material is desired.

In reality a mixed form will be applied by every archive, because important data that are offered should not be refused and probably there will not be enough money to immediately do all categorizations and transformations. In the DOBES programme we can differentiate between material that is created as part of the documentation work and other material. For the first kind of material a more restrictive policy close to the "immediate model" was chosen.

[11] An excellent example for the "immediate way" is the CHILDES corpus created at the CMU [8].

5. Access Issues

With respect to accessing the archive material we have to distinguish between two aspects:

- The technology that gives access to the material.
- The management of rights that allows individuals or groups to access the material.

5.1 Access Technology

In the DOBES project much attention was given to technologies that allow the archivist and teams to create and maintain a properly organized corpus and to access the different data types, in particular the complex annotated multimedia recordings that include sound and/or video. Recently, we developed new ways for supporting general users who might be interested in the DOBES material.

As already mentioned all metadata is created according to the IMDI framework, i.e. the DOBES domain of resources is described by interlinked and open accessible XML-files. A browser was built that allows navigation in this domain (see figure 4). It offers possibilities to browse in this domain and to execute structured searches such as “give me all resources spoken in Jaminjung by a 60 year old female speaker”. This browser can easily be downloaded and installed on any modern desktop computer. It can be seen as a shell provided by the archivist that is not mandatory to be used.

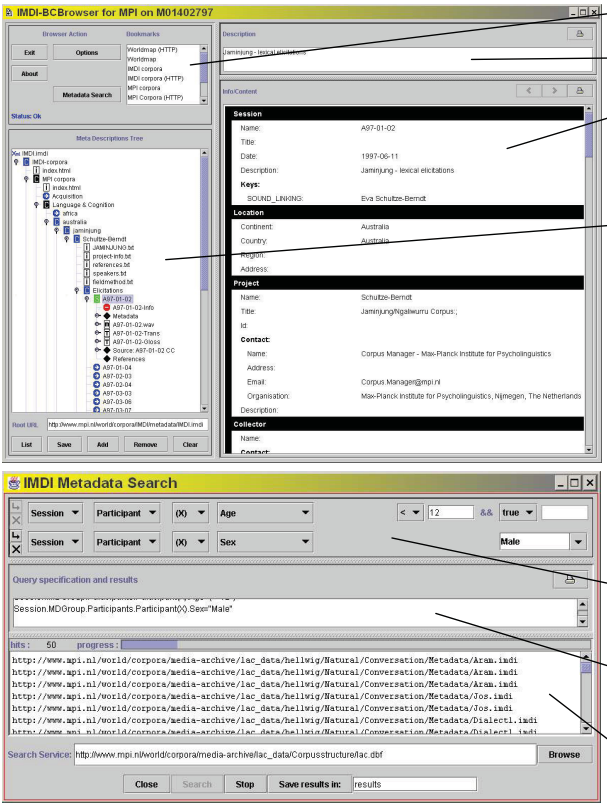


Figure 4 shows the user interfaces of the IMDI browser and the integrated search component that is part of the browser.

All metadata files are available via the Internet. They can be accessed via the HTTP server by sending normal requests allowing everyone interested to create his own shell. In fact the archivist just completed alternative methods to navigate in the DOBES corpus. It is now possible to use a normal web-browser such as Internet-Explorer to browse through the linked metadata descriptions. While browsing, the XML-files are transformed on the fly via XSLT scripts to HTML and the path is stored so that it is easy for the user to go back into the hierarchy. Also a Google-type full-text search[12] is provided that allows users unfamiliar with the details of the IMDI set to carry out simple searches on the metadata

[12] Of course, databases and index files are created in the background to support efficient search. Yet we did not make an analysis that can compare the precision and recall quality of the two methods for the DOBES corpus: structured search on structured data versus unstructured search.

descriptions including the content of the unconstrained fields in the IMDI set. This new shell was produced, since we had to accept that many users find it too complicated to download new tools, to install and use them.

In both methods for resource discovery it is possible to immediately start looking at a resource once it was found. Within the IMDI browser it is possible to create a configuration file that contains the tools one would like to use when individual or bundle of files occur. This allows the user to start tools that can operate on complex annotated recordings. Using the normal web-browsers one can only select an individual resource such as an MPEG file and start the preferred video player. First tests were made to use SMIL [9] to visualize annotated media recordings.

To guarantee a coherent metadata domain a professional editor was built that supports controlled vocabularies and constraints for the corresponding metadata elements. Also a tree-building tool was developed that allows users and managers to easily create linked structures.

The DOBES metadata descriptions are also offered to OLAC[13]/DC[14] [10,11] service providers, since a mapping from IMDI to OLAC concepts exists and the OAI harvesting protocol[15] [12] is supported. Of course, it is up to service providers to design their user interfaces and the scope of their services.

Another tool provided by the archivist is a multimedia annotation and exploitation tool called ELAN (see figure 5). It allows the user to create complex annotations on audio and/or video recordings and to visualize and analyze them. Again the users are free to choose whatever tool they like for exploiting the data in the archive, since all recordings and all annotations are accessible via the HTTP server (given proper access rights) as individual files and stored in well-documented open formats. The annotations are encoded as XML-files according to the EAF schema [13], the videos in MPEGx and the audios as WAV files. The archivist tries to offer a shell that can be used for exploitation as easily and effectively as possible.

It should also be added here that the workflow system is supported by a specially designed database, which provides the archive managers with information about the exact status of each of the many transactions. This database is coupled with the metadata

[13] The Open Language Archives Community has defined another metadata set with a more general coverage.

[14] The Dublin Core Initiative has created a metadata set that is meant to be used for all sort of web resources.

[15] The Open Archives Initiative has developed a simple protocol that allows service providers to harvest metadata records from data providers.

descriptions since some administrational data is contained in them [16]. The documentation teams can access this database too to check the status of their material.

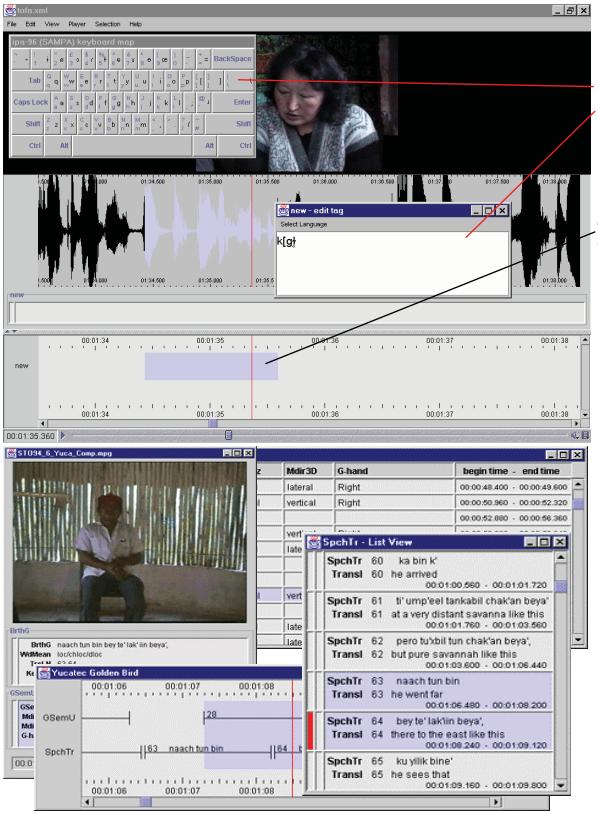


Figure 5 shows typical screenshots of the ELAN annotation and exploitation tool. At the left a video is shown together with the audio information. A segment has been selected and can now be annotated in a number of writing systems such as IPA, Chinese, Hebrew etc. At the right side it is indicated that ELAN supports a number of different views on the data and all viewers are synchronized.

5.2 Access Rights Issues

The purpose of the DOBES archive is to offer as much of the material as possible to interested users. However, there are many reasons that will not allow us to make all material openly available. First, ethical and legal aspects have to be taken into account.

[16] In the IMDI set it was chosen to not burden the set with workflow type of information. This would increase the number of elements most of which are irrelevant for users.

Some speakers from indigenous communities may not want certain recordings or certain texts to appear open in the web. This may have religious, political or other important reasons. Second, the work of young researchers creating linguistic annotations and working on dissertation projects should be protected for some time by allowing only selected persons to access their data. Thus, there may be many reasons to limit access to the archive resources. Different types of resources will have to be treated differently and it is obvious that access rights will change over time[17].

In DOBES the researchers are responsible for defining access rights and policies. They

- know the speech communities and the consultants
- know the rules and dangers of our industrialized world
- are the best to mediate between the community wishes and formal rights management

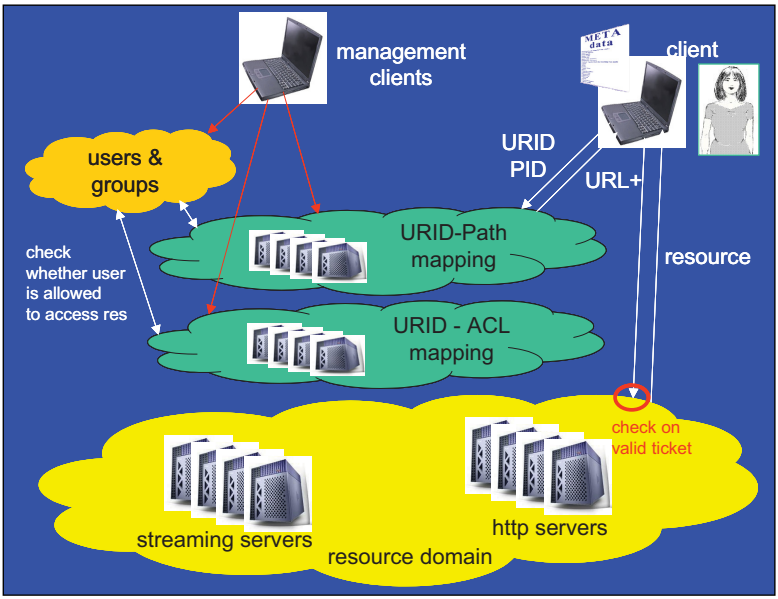


Figure 6 indicates a possible access rights management system that includes its three essential pillars: (1) A mechanism to define users and groups, (2) a mechanism that allows administering and resolving unique resource identifiers and (3) a mechanism that allows to associate access rights with URIDs.

[17] According to McConvell (AIATSIS) defining access rights is a matter of continuous discussions with the indigenous communities.

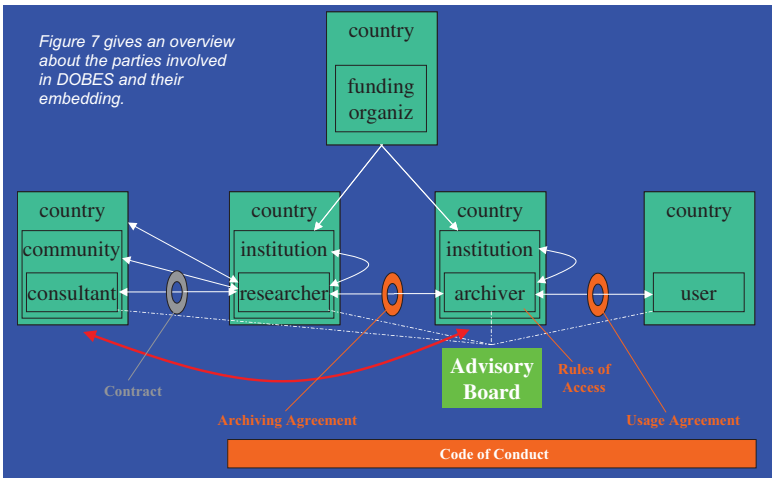
Given the amount of resources and the potentially large group of interested users, an efficient access rights management system has to be developed. Since such a system is not yet available, simple procedures are applied now. The consequence of this is that only few resources are open for everyone.

The intention is to establish a system that allows managing access rights such that at the end of the documentation work of the first teams the researchers come up with detailed access rights. In order to not create too heavy a load on the DOBES archive managers and to make best use of the knowledge of the researchers the system has to support delegation mechanisms. The system has to include its own user administration that must also allow defining user groups. In addition, we have to anticipate that resources will be available from various sites, i.e. there will be copies of the resources. Therefore, it will be necessary to introduce unique resource identifiers and associate such a URID with each resource. Access rights have to be linked to the URID and the metadata has to point to the URID. A resolving mechanism has to map URIDs to physical paths as described in figure 6.

Currently, the archivist is investigating the possibilities to implement such a system. It has to be designed and implemented with great care to satisfy efficiency criteria and to make sure that it is robust and safe.

6. Ethical and Legal Issues

The previous chapter has shown that access rights issues are very sensitive matters. It was understood that in the DOBES project many conflicts could arise due to the different parties involved. Figure 7 gives an overview of the complex situation we are faced with.



The parties involved are the consultants from the various speech communities, the researchers of the different teams[18], the archivist, different groups of users of the material and finally the funding organization. All are embedded in communities and institutions that have their legal and ethical regulations. In addition, since DOBES is active on all continents all of these communities and institutions are embedded in different national regulations. So from a juridical point of view the situation is extremely complex.

Moreover, possible ethical conflicts are even more difficult to handle. Therefore, the DOBES programme agreed on a number of principles that are openly documented and seen as guidelines for the work of everyone. A Code of Conduct is the central document here, since it specifies the rules that have to be adhered by everyone. Further, where necessary and applicable, contracts should be made between the consultants, their communities or their representing organizations that specify the status of the material. An Archiving Agreement between the teams and the archivist defines the rules according to which both parties will interact with each other. The archivist has created a document that specifies the rules for archiving. Another document will have to be prepared that specifies the rules for using the material by users that have got access permits.

Since all these documents will not prevent that problems may arise and since the responsible researchers that could take decisions will not be accessible anymore after some point in time, we need a forum that can take care of any conflicts arising. Therefore, the DOBES programme established an Advisory Board with well-known experts in fieldwork who know the situation in the various regions to help solving conflicts.

7. Training Issues

In the pilot phase many teams were interested in training courses about various aspects involved in the documentation of languages, in particular to learn about modern tools and procedures. Therefore DOBES organized comprehensive 5-day training courses with theoretical and practical sessions after new teams have been given grants. The training courses covered the following topics:

- The transfer of the linguistic agreements to new teams.
- The in-depth explanation of the different workflow options.
- The discussion of practical fieldwork issues such as optimal power management, how to do good video and audio recordings, how to handle large data quantities in the field, how to operate with audio and video signals on notebooks etc.

[18] In some cases German hosting institutions for documentation teams make the situation even more complex.

- The explanation of the IMDI framework to create a well-organized metadata described corpus and do a hands-on training about how to use the IMDI-editor and browser.
- The explanation of how to use state-of-the-art tools such as for creating multimedia annotations, lexicons and other linguistic data types including hands-on practice.
- The explanation of how to do speech analysis with state-of-the-art tools including hands-on practice.

These training courses were absolutely necessary to create the positive atmosphere and level of interaction needed to come to the coherent archive described above. The content and structure of the training courses need to be adapted continuously to cope with the most recent development in technology and with the changing expectations of the teams.

8. Conclusions

The DOBES programme has been a quite successful programme so far in particular when considering its relatively complex construction. Within shortest time it was possible to agree on linguistic and technical frameworks for starting the documentation work on a high level. While the archivist has primarily been focusing on technologies and procedures that address the proper organization and coherence of the archive, it is now time to shift the focus on questions of how to use the archive. Next year the first teams will finish their documentation work. The archivist will then have a complete set of relevant documentation resources. This will be the moment to summarize and discuss the experiences. It is intended to organize a conference in 2004 where these experiences will be presented and discussed.

Many topics could only be briefly touched on in this paper. For more detailed information we would like to refer to the following web sites: www.mpi.nl/DOBES, www.mpi.nl/IMDI, www.mpi.nl/tools.

We would like to thank all the teams from the DOBES programme in particular those that participated in the pilot phase for the sometimes difficult, but always encouraging discussions and for the high level of collaboration.

References

[1] Advanced Glossing www.mpi.nl/DOBES/applicants/Advanced-Glossing1.pdf
[2] Shoebox www.sil.org/computing/shoebox/
[3] Transcriber www.etca.fr/CTA/gip/Projets/Transcriber/
[4] Praat www.fon.hum.uva.nl/praat/
[5] ELAN: www.mpi.nl/tools
[6] International Phonetic Font www2.arts.gla.ac.uk/IPA/fullchart.html
[7] IMDI framework www.mpi.nl/ISLE; www.mpi.nl/IMDI
[8] CHILDES childes.psy.cmu.edu/

[9] SMIL www.w3.org/TR/smil20/
[10] OLAC www.language-archives.org/
[11] Dublin Core dublincore.org/
[12] OAI www.openarchives.org/
[13] ELAN Annotation Format www.mpi.nl/tools