

# Methods of Language Documentation in the DOBES project

P. Wittenburg, U. Mosel<sup>+</sup>, A. Dwyer<sup>++</sup>

Max-Planck-Institute for Psycholinguistics

<sup>+</sup>University of Kiel

<sup>++</sup>University of Kansas

Contact: peter.wittenburg@mpi.nl

## Abstract

The DOBES program for the documentation of endangered languages, started in September 2000, has just completed its pilot phase. Eight documentation teams and one archiving team worked out agreements on formats, tools, naming conventions, and encoding, especially the linguistic level of encoding. These standards will form the basis for a five-year main phase, which will include about 20 teams. In the pilot phase, strategies to set up an online archive incorporating redundancy and regular backup were developed and implemented. Ethical and legal aspects of the archiving process were discussed and amounted to a number of documents to which all participants have to adhere to. Tools and converters developed within the pilot phase are available to others.

of audio and video recordings. This paper provides an overview of the DOBES framework as it has been developed during the pilot phase. For details, see the DOBES website ([www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)).

## 1. Introduction

The Volkswagen Foundation-sponsored DOBES program<sup>1</sup>, supporting the documentation of endangered languages, was launched in September 2000. Since two-thirds of the world's 6000 languages will have died out by the end of the 21<sup>st</sup> century, linguistic and cultural documentation has become a most urgent task.

The documentation of a language as understood by the DOBES program is in many respects innovative. While earlier language descriptions usually consisted of a grammar, a dictionary, and a collection of narrative texts in print, recent language documentation such as the DOBES project are centered around audio and video recordings of different speech situations with sound-linked transcriptions, translations and commentary. The aim is to allow future generations to reconstruct how the language was used in various social interactions and how it encoded the traditions and cultural values of the speech community.

To achieve these objectives, the DOBES teams recommended that language documentation should:

- be based on multimedia recordings;
- be as theory-neutral as possible;
- be useful for many disciplines, for the interested public, and the speech community;
- be presented in such a way that they can be understood by researchers of other disciplines without any prior knowledge of the language in question.

The one-year pilot phase included eight documentation projects and one digital multimedia archive project; their aim, in addition to documentation, was to work out the linguistic, technical, and ethical framework of language documentation. The languages were: Awetí, Trumai, Kuikuro (all in Brazil), Wichita (US), Tofa (SU), Salar, Monguor (China), Teop (Papua New Guinea), and Ega (Ivory Coast). The tasks of the archive project included defining policies and the flow of data between the documentation projects and the archive, and developing efficient tools for the annotation, glossing, and translation

## 2. Linguistic Issues

### 2.1. Typological differences

The DOBES program covers languages of the typologically greatest diversity, which demands a high level of flexibility of the archive structure and has direct consequences for data architecture design. For example lexical databases need to facilitate search functions for affixes, stems and roots of polysynthetic languages. Furthermore, the native speakers of such languages might wish for different search functions than linguists as the DOBES Wichita project has pointed out. Since all projects work in close cooperation with the speech communities, it goes without saying that the archive feels obliged to accommodate the needs of the indigenous speech communities.

### 2.2. Data Types

- The pilot phase teams decided that a documentation should include the following types of data as a minimal requirement: a brief description of the genetic affiliation of the language and its prominent typological features;
- an outline of the sociolinguistic context (e.g. the number and distribution of speakers, degree of multilingualism in the speech community, educational system), the research history, and the circumstances of the documentation;
- an annotated corpus of audio recordings and, where possible, also video recordings of different language genres (e.g. myths, anecdotes, procedural texts, casual conversations, political debates, and ritual speech events) accompanied by a transcription, a translation, and content and linguistic commentary as needed;
- a Metadata Description for each recorded session which in a standardised format provides

<sup>1</sup>Dokumentation der bedrohten Sprachen,  
[www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)

Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

information on when, where, and by whom the recording was made, which equipment was used, who the speakers were, etc.;

- a detailed phonetic/phonological description, including an inventory of segmental and prosodic phonemic distinctions and allophonic/allotonic rules;
- a description of the transcription conventions (see below) and also, where applicable, a description of orthographic conventions; a list and explanation of all abbreviations and symbols used in the documentation;
- references to previous studies on the language and culture.

The corpus of recordings, together with their transcriptions and translations, constitute the main body of the documentation. In addition, a documentation can include a grammatical sketch, a bilingual dictionary, and specialized descriptive linguistic or ethnographic essays.

Further, some projects may add data elicited in interviews with native speakers. For example:

- a collection of recorded and transcribed data on phonetic/phonological phenomena which cannot be worked out on the basis of text recordings and are not covered in the required description of the phonetic/phonological system;
- a collection of morphological forms which are presumably not completely attested in the corpus (e.g. inflexion paradigms, derivational morphology) and other data on the grammar of the language (such as elicited sentences) which could serve a future analysis of grammatical phenomena;
- lexical data: words of certain semantic fields, nomenclature (e.g. plant and animal names, kinship terms), numerals, idioms and proverbs;
- photographs and drawings, which, for instance, illustrate language data on ethnobotany, the natural environment, or artefacts of the material culture;
- music recordings;
- videos of cultural activities (e.g. dances, house-building);
- other systematically-collected data which are relevant for the understanding of linguistic phenomena as part of an extensive socio-cultural complex.

## 2.3. Recordings

The basis of all documentation work are the recordings of the language. Wherever culturally acceptable, videos should be recorded simultaneously to audio recordings, since facial expressions, gestures and body postures play an important role in communication. Video recordings can document cultural practices such as dances, games, and farming. Technically, the importation, conversion, and annotation of the video stream is approaching the ease of audio-recording processing. In order to be considered complete, a language documentation should consist of a considerable number of linguistically-annotated recordings, analyses, and supporting materials. This demands efficient annotation methods and tools, in order to make maximum use of time for documentation.

## 2.4. Annotations

### 2.4.1. Tiers

The DOBES teams agreed on a minimal annotation tier standard mandatory for any transcribed data submitted to the archive<sup>2</sup>.

We established two obligatory tiers: *rendered text* and *gloss*. The *rendered text tier* is a form of the original A/V stream rendered in a segmental transcription or orthography chosen according to individual project needs. IPA (the International Phonetic Alphabet) is required except in certain cases. The *gloss* (“translation”) tier is simply free translation of the Rendered Text tier into a major language.

These obligatory minimal tiers do not constrain an individual project from setting its number of tiers higher. Furthermore, there may be several types of one tier (for example, a glossing tier in the local lingua franca and another in English.)

Each team can thus develop its own appropriate annotation model.

One maximally extensible annotation model that has emerged from the pilot phase is *Advanced Glossing* (Lieb and Drude [1]). This is a highly detailed scheme for up to 24 annotation tiers for morphology and syntax, which clearly distinguish formal tiers from functional ones.

The tradeoff of the chosen methods is an end product of a few, densely annotated materials or many, lightly annotated materials. The choice depends on the aim of project. In general, a mix is advisable.

### 2.4.2. Tagsets

Nearly all linguists annotate their data for morphological, syntactic, prosodic, or other information, and they do it in a fairly idiosyncratic way. The EUROTYP proposal [2] was the first attempt to develop a standard for such annotation. However, it was designed for European languages only, and -- as the first proposal of its kind -- did not have the input of a number of specialists. DOBES took up the problem of morphosyntactic annotation with particular attention to (1) clarifying formal from functional notions, (2) keeping linguistic levels separate in annotation, and (3) developing

---

<sup>2</sup> A. Dwyer Nov. 2000 *DOBES linguistic markup scheme: Towards a Minimal Annotation Standard for Encoding Linguistic Information*. DOBES technical paper.

groundwork for a tagset for *non*-European languages. The beginnings of an ontological hierarchy of tags, presented to E-MELD by Dwyer in 2001, stimulated the formation of an annotation group also within E-MELD [3]. At present, there is broad consensus on the common use of generic labels (which are now in a revised hierarchy); these would, however, have to be defined (assigned attributes) by each individual project.

This approach affords maximal flexibility based on individual project aims, while creating a minimal standard for annotation.

## 2.5. Lexica

Most, if not all DOBES projects will include a lexicon of some kind into their documentation (lexicon understood here as a cover term for all kinds of lexicographic work ranging from simple alphabetic wordlists with translation to dictionaries with entries of a complex microstructure. The dictionaries also vary with regard to their macrostructure. While some prefer strict alphabetization, other choose nesting where derivations are found under the word they are derived from, or they organize the head words in the form of a thesaurus. Which kind of micro- and macrostructure the individual projects choose depends on numerous factors such as the structure of the language, the lexicographic tradition of the dominant language of the region, the preferences of the speech community etc.

Since the time frame of the documentation projects does not allow to plan comprehensive dictionaries, the projects will produce corpus based dictionaries, dictionaries which only cover limited subject areas (e.g. kinship terms, terms for body parts, animals, plants, etc) or a combination of a corpus based dictionary and a thematic dictionary [4].

A recent study [5] has shown that there are large differences as to what type of lexicon has to be constructed, and which attributes it should have. Due to the differences in requirements of the different teams, no uniform structure or attribute set was agreed on. However, it was recognized that documenters need flexible lexica which can be easily adapted to their needs, and which can be shared over the Web. Such a shared lexicon was designed and implemented [6].

## 3. Archiving Issues

### 3.1. Task of an Archive

The first and most important question which had to be addressed in the pilot phase with respect to the archive was the question of what an Archive of Endangered Languages will constitute. It was agreed that an archive has to be a facility which stores the documentation material such that it can be accessed for a long time, that it is available as a coherent set of data types which can be interpreted even years later and that access to it has to occur via Internet, i.e. the archive has to be an online archive. This definition does not include an open access to all resources, however, the information about the material, the metadata, should be openly available.

The offering of an archive depends on its potential users. To answer the question which the users are is difficult, since the archive is not primarily intended for people living now, but for future generations. Nevertheless, we can identify a set of groups which may

have an interest in the material. Besides the linguists, ethnologists and other researchers we see interests from school and university educators, journalists, and especially from the indigenous people themselves. All users have completely different requirements. When the archive wants to serve these needs it has to come to a neutral representation and presentation of the data. The creation of a guided web-site where some material is presented for to achieve attraction can only be done by the researchers involved. The wishes of the indigenous people themselves have to be taken very serious. But again the researchers are the central anchor points to present the material such that they can be of use for these communities for example in school education.

The archive definition raises a number of problems which will be discussed subsequently.

### 3.2. Long-term Storage and Standards

The term “archive” implies that the archived material will be stored for ever. On the other side we have multimedia material and we have the duty to have the data available online. Only information technology allows us to serve these requirements. However, in information technology we are hardly able to guarantee a long lifetime of data, if we look to a single storage medium such as CD-ROM. For the medium itself lifetimes of about 30 years are reported. But technology development is so rapid that a given technology may after a decade be no longer available. We are confronted with a paradox, since we want to guarantee long-term data storage at a time when lifetimes of the storage media are increasingly shorter.

The solution is an organizational one. The archiving team decided to use tape technology as principle storage medium although disks are getting more and more attractive to store large data volumes. In the center is a Hierarchical Storage Management System (HSM) which is used to generate automatically two copies at distinct locations. HSM will guarantee that a number of files will be available on disk caches to guarantee fast access times. The fact that a certain tape technology such as AIT 3 which was selected to be the base for the DOBES material will only survive for about 5 years, leads to the need to copy all data at regular intervals to new tape media and to replace the robot if it cannot handle various tape formats. These operations can only be done automatically. So a continuous effort is required to fulfill the claim of “eternal data storage”. Worse, we cannot give guarantees, since the survival of the data will be dependent on future generations and future managements. With paper as storage medium this was not principally different. However, good paper could be stored for hundreds of years without that work had to be invested.

The problem of making the data available after many years is strongly linked with the usage of open standards. Given that the archive has solved the storage problem, still the problem of interpreting the data correctly has to be solved. Here we are confronted with two levels: (1) The formal encoding level and (2) the linguistic encoding level (this aspect was discussed above). The formal encoding level has to do with adhering to open standards as extensively as possible. The DOBES program decided to follow the following standards:

- MPEG1/2 for video encoding [7], since its principles are openly documented. MPEG2 will be

used for the archive and where enough bandwidth is available. MPEG1 will be used for Internet-based activities. Several file formats are accepted such as MPG, AVI and QT, since conversion can be done easily.

- WAV format [8] and PCM encoding (at 44/48 kHz) for speech files. Here it has to be analyzed carefully whether a compressed format such as MP3 can be accepted (see below). Of course, MP3 would reduce the amount of storage capacity needed considerably.
- XML (eXtensible Markup Language) as syntactic and structural base for all type of textual material. A number of less dominant data types such as sketch grammars are typically provided as WORD files and there is yet no clear structure. Currently, these files are represented as HTML files. For metadata and annotations XML-based schemas have already been worked out. For lexica this still has to be done.
- UNICODE is an important step, since the bad practice will hopefully find an end that individual researchers define their own fonts (including some individual character representations which are not documented). Currently, all incoming material is mapped to UNICODE characters. However, not all wishes can be satisfied, i.e. there are glyphs which are not yet included in the UNICODE set. It is badly received that the User-Definable Region in UNICODE is practically not available to projects, since there is the evidence that Microsoft and Adobe are already using this space. It is clear that we need an extension of the UTF-8 standard to add for example IPA characters and tone indicators not yet represented by UNICODE.

Currently, some work is going on to investigate whether compressed audio formats such as MiniDisk (MD) and MP3 (the audio compression defined for MPEG2) can be used for archiving purposes. It can be expected that both formats will be replaced after a number of years. While MD is associated with devices from specific companies, MP3 is a widely accepted and well-documented format. Therefore similar to MPEG2 there would be no principle reasons to not use MP3 for archiving purposes. Another important question which was addressed and which is investigated at this moment is whether the audio representations created can be analyzed following the well-known speech analysis algorithms such as the calculation of the pitch contour or of the tube configuration simulating the vocal tract via reflection coefficients. First results seem to show that the implicit filtering done by MP3 does not lead to essential analysis errors.

Guaranteeing conformity by the archivist within DOBES cost much work in the pilot phase. This is due to the fact that we lack good and robust tools which support these standards. Within DOBES and other related projects essentially two tools have been developed which give this support. For the metadata the IMDI schema [9] was used and for annotations the EUDICO Annotation Format (EAF) [10].

However, in the community it is still an open discussion in how far an archive should collect data in whatever format it is delivered or whether strict standards

have to be followed. We feel that the pure collection without standardization on open formats will create an unsolvable management task and access problems. It must be doubted whether such a liberal approach can be successful in the long run.

### 3.3. Ways of Access

Different ways to access the DOBES data were discussed. Due to the contracts all documentation teams are requested to provide the data to the archivist. Nevertheless, the individual teams can of course store their data and present it to the public. The way they do it is in the responsibility of the teams. The individual teams don't have to give guarantees.

The archivist has to define the ways to access the stored data. The legal and ethical aspects will be discussed later. First, the level of searching or browsing in a catalogue is relevant so that a user can find out which data is available in the archive. Here the IMDI framework is used. All major resources are described by metadata following the IMDI standard which was strongly influenced by members of the DOBES team. Using the BCBrowser [11] each user can browse in a hierarchical metadata domain which also integrates data types such as lexica or sketch grammars at the appropriate places. The browser also allows to search in such metadata domains. In both cases the user can directly start tools when resources were found. Of course, it is possible to read the XML-based IMDI files with other programs. In future we expect other services which will make it possible to interpret the IMDI files and build services on top of it.

With respect to the annotated multimedia resources the user also has two possibilities (if he has the rights of accessing them): (1) He could use the EUDICO tool set [12] to visualize the media files synchronized with the annotations. (2) He could use his own tool to either play the media files, view the annotations which are in XML-based files the structure of which is described by the EAF Schema. The lexica are currently not standardized and are delivered in various formats which makes access to them more difficult. The archivist has built a first version of a lexicon tool, but this is not yet generic enough to represent all the different structures. Other less important resources are mostly converted to HTML files which are associated with a node in the metadata hierarchy and rendered by the browser.

So, for the metadata and the annotated recordings the archivist offers tools which gives many flexibility and comfort to access the main resources. But the user can choose his own ways. To support the indigenous people the archivist is currently busy to integrate components which allow to print resources or segments of them.

### 3.4. Workflow and Data Management

It turned out to be of primary importance to come to agreements between each documentation team and the archivist about the interaction. Each workflow scheme describes the kinds of resources that can be expected from the documenters and in which formats they will be delivered. In the pilot phase it was agreed that the media digitization was mainly done by the archivist to assure format coherence. All steps to come from field recordings to the metadata described annotated resources were described in detail. In the pilot phase it turned out to be

very helpful for most of the teams to describe these steps such as how to split media files and to describe them correspondingly. Also conventions for the naming of files are part of the negotiations. All these agreements were necessary to manage the data flow from the 8 documentation teams.

On the archivists side data management is of great relevance, since media files, annotations, series of photos, derived data such as lexica and metadata files are delivered at different moments in time. However, they are related in specific ways and these relations have to be conserved to come to a coherent corpus. The metadata descriptions play the central role here, since they organize the corpus and establish the relationships between the various data types. Workflow databases were used to document the states of the different activities at the archivist side to handle the complex logistics. In the main phase we expect up to 20 documentation teams which all will deliver data at completely random moments. This flow has to be efficiently managed.

### 3.5. Tools for Data Creation and Conversion

It was not surprising that many teams had already data which was made with some tool and existed in some format - partly incompletely documented. The conversion of this legacy data took much time for the archivist, since the documentation teams often lack the knowledge or don't have the manpower. In the pilot phase a number of tools were discussed and supported to allow the teams to carry out their work. The following tools were accepted for new data creation, partly since the teams were already used to these tools and since there was no alternative:

- BCEditor [13] for the creation of standardized IMDI metadata descriptions. Some WORD templates were also used initially, but they required much manual control and correction work.
- MediaTagger [14] for the creation of multimedia annotations.
- Transcriber [15] to create annotations of audio files.
- SHOEBOX [16] to create annotations and lexica.
- MS WORD to create annotations and lexica.
- MS EXCEL spreadsheets with metadata parts.
- A specially designed relational database in FoxPro to cover a lexicon.
- PRAAT [17] to do speech analysis and annotations.

To deal with these different formats a number of converters were written, although much handwork remained to be done to meet the goals;

- ECONV - to convert between SHOEBOX, Transcriber and EAF.
- WORD2EAF - a converter between structured WORD files and EAF, where the user has a simple language to describe the structure of his word file.
- WORD2PRAAT - to convert between WORD and PRAAT.
- Some XSLT scripts to convert between versions of XML files.

Most of these converters dealt with the problem of character conversion by supporting mapping tables.

It was clear that especially with the perspective of having about 20 documentation teams in the main phase the format and encoding variety with respect to new recordings had to be reduced to make the archiving task tractable. For the main phase the following file formats for media and annotation files will be accepted:

- MPEG1/2 for video files<sup>3</sup>
- WAV for audio files
- EAF<sup>4</sup> for the annotations of multimedia files (sound and video)
- SHOEBOX format for annotations and lexica
- IMDI for metadata

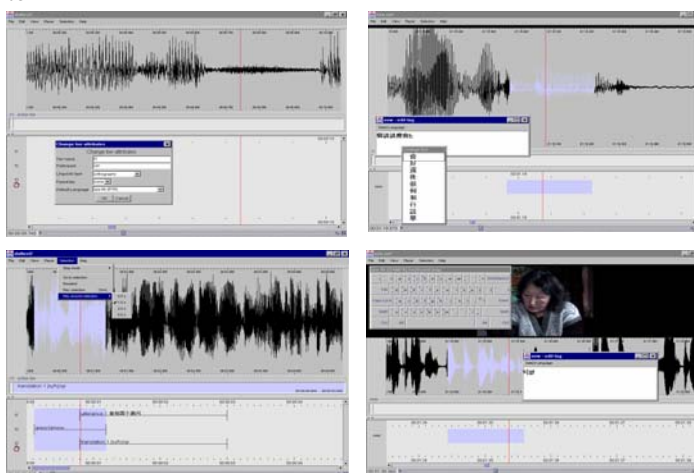


Figure 1 shows some screen shots from the EUDICO Annotation Tool (ELAN) which can be used both for the annotation of audio and video files. It provides input methods for a number of character sets under which is Chinese, supports UNICODE and has rendering engines for a number of complex writing systems such as Arabic and Bengali.

This very restricted set of formats will ensure that the archiving task is manageable. The user has to make sure that at least for new recordings the mentioned formats are created. Of course, this can only be required when the archivist provides tools which are able to create these formats. Essentially these tools are recommended:

- BCEditor to create metadata descriptions
- ELAN (EUDICO Annotation Tool) to create annotations for audio and video files
- SHOEBOX to create annotations and lexica
- WORD to create lexica and other data types

Each team can define their own set of tools, but then it is their task to show format compliance which in general is beyond their skills. So the usage of the suggested software tools is highly recommended. In a few years of time we expect that there will be more tools which support formats such as Atlas Interchange Format [18] or EAF.

<sup>3</sup> The usability of MPEG4 will be checked as well.

<sup>4</sup> The Atlas Interchange Format (AIF) will also be supported when it is complete and robust enough for the DOBES purposes.

### 3.6. Field Situation

The field work of the DOBES documentation teams is often carried out under very bad conditions. Therefore, the work to be carried out in the field and the equipment used has to be checked carefully. On the other hand teams often go for a few months to their field research site, i.e. the possibilities have to be checked in how far some of the work can already be done in the field. Another consideration is that increasingly often it turns out that the indigenous people are eager to work themselves with the new multimedia software tools and to create for example transcriptions.

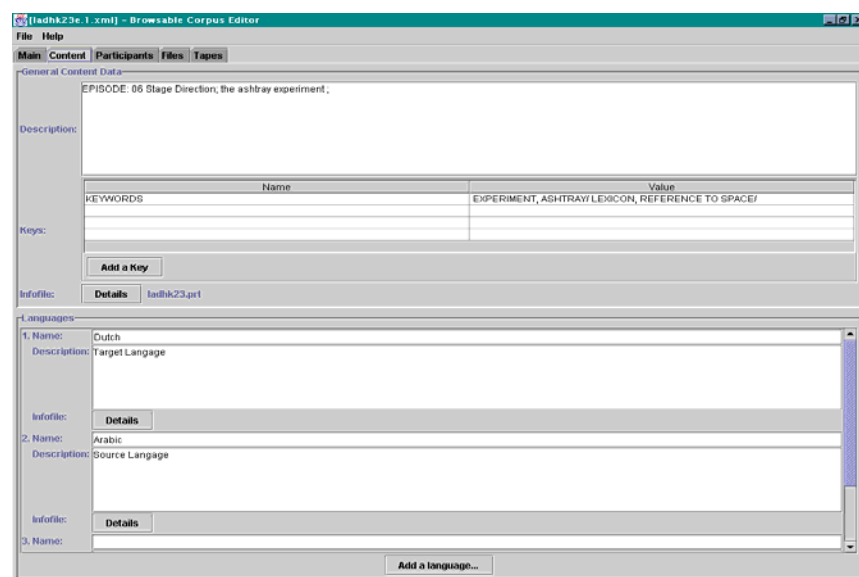


Figure 2 shows a screen shot from the BCEditor which supports controlled vocabularies and constraints to achieve a high quality of the generated metadata descriptions.

Currently, tests are carried out to define environments for video digitization in the field such that the teams being in the field for a few months can already start annotating their videos in the field. A number of compatibility checks have to be carried out, the conversion from Digital Video (DV) to MPEG1 or MPEG2 has to be sorted out and the storage problem has to be solved in field circumstances. Good recommendations would help some teams.

### 4. Ethical and Legal Issues

The trend towards online archives of multimedia material creates new challenges with respect to the proper treatment of legal and ethical aspects. Of course, the indigenous people as well as the researchers have rights with respect to the data stored by the archive. In addition to this the recorded people have the right to determine whether their faces and voices may appear openly on the Web. In the DOBES internal discussions an analysis was made about the parties involved and their rights and interests. The juridical situation was recognized as being very complex, since the laws of several countries are involved and since the indigenous people - the individuals or the community - have also basic rights.

Based on this analysis, the discussion with juridical experts and experienced field researchers a few documents were developed which describe the duties and rights of all the parties involved. Most important is a code of conduct which guides all parties in their activities. The code of

conduct especially requires an ethically correct behavior. Within the DOBES program there is a clear trend to a greater openness of the material, but also the researchers work can be protected within the first 3 years. Further, documents describe the relation between archivist and documentation teams on the one hand and archivist and users on the other hand.

Due to its construction the DOBES archivist does not have direct relations with the indigenous community. Experienced field workers have correctly pointed out that one can expect conflict situations which question the behavior of the archive. Since the archivist itself does not have the competence to decide in such situations, it was necessary to build an Advisory Board. This AB was built and contains a mix of experienced field workers being active on the various continents.

### 5. Links to other Initiatives

Recently, a number of initiatives were started which have similar tasks. To be mentioned here are especially the ASED A project in Australia [19], the E-MELD and AILLA [20] projects in the US and the LACITO project [21] in France. Of course it is very interesting to achieve a high degree of interoperability on various levels with those initiatives. Common workshops document the will to accomplish this.

The DOBES archive will be integrated in the new Integrated European Resource Area (INTERA<sup>5</sup>) project which will be funded by the EC and which has as goal to create an IMDI based, browsable and searchable metadata domain in Europe and beyond.

Other collaborations such as for example to come to common agreements on XML-based annotation formats were already mentioned.

### 6. State of the Archive

Despite all the difficulties and the discussions necessary in the pilot phase the documentation teams have produced already much data. These are especially multimedia recordings which were analyzed and annotated according to the various agreements. Each team also developed other types such as lexica, grammar descriptions, field notes, annotated series of photos and much more. These were all integrated into a browsable and searchable metadata domain for DOBES. This metadata domain is open for the users and also some resources are freely available. From the DOBES web-site tools can be downloaded which allow to operate on the data. However, also the users have to agree to behave according to the requirements of the code of conduct.

### 7. References

[1] H. Lieb, S. Drude (2000) Advanced Glossing: A language documentation format. Unpublished DOBES working paper.

<sup>5</sup> New EC funded project to build up an European language resource area organized by IMDI type metadata.

- [2] Eurotyp: [www.linguistlist.org/issues/10/10-1099.html](http://www.linguistlist.org/issues/10/10-1099.html)
- [3] E-Meld: [listserv.linguistlist.org/archives/e-meld.html](http://listserv.linguistlist.org/archives/e-meld.html)
- [4] U. Mosel (2002) Dictionary making in endangered speech communities. In Proceedings of the LREC Preconference Workshop on Tools and Resources in Fieldlinguistics (to appear)
- [5] W. Peters, P. Wittenburg, S. Drude (2002). Analysis of Lexical Structures: Towards an Abstract Lexicon Model. In Proceedings of the LREC 2002 Conference (to appear)
- [6] D. Harrison, G. Gulrajani, P. Wittenburg (2002). SHAWEL: a sharable web-based lexicon. In Proceedings of the LREC Preconference Workshop on Tools and Resources in Fieldlinguistics (to appear)
- [7] MPEG: [mpeg.cselt.it](http://mpeg.cselt.it)
- [8] WAVE format: [ccrma-www.stanford.edu/courses/422/projects/WaveFormat](http://ccrma-www.stanford.edu/courses/422/projects/WaveFormat)
- [9] IMDI Metadata Schema; [www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)
- [10] EAF Schema: [www.mpi.nl/tools](http://www.mpi.nl/tools)
- [11] BCBrowser: [www.mpi.nl/tools](http://www.mpi.nl/tools)
- [12] EUDICO/ELAN Tool: [www.mpi.nl/tools](http://www.mpi.nl/tools)
- [13] BCEditor: [www.mpi.nl/tools](http://www.mpi.nl/tools)
- [14] MediaTagger: [www.mpi.nl/world/tg/mt/mt.html](http://www.mpi.nl/world/tg/mt/mt.html)
- [15] Transcriber:  
[www.etca.fr/CTA/gip/Projects/Transcriber](http://www.etca.fr/CTA/gip/Projects/Transcriber)
- [16] Shoebox: [www.sil.org/computing/shoebox](http://www.sil.org/computing/shoebox)
- [17] PRAAT: [fonsg3.let.uva.nl/praat/](http://fonsg3.let.uva.nl/praat/)
- [18] Atlas Interchange Format: [www.nist.gov/speech/atlas](http://www.nist.gov/speech/atlas)
- [19] ASEDA:  
[www.aiatsis.gov.au/rsrch/rsrch\\_pp/ased\\_abt.htm](http://www.aiatsis.gov.au/rsrch/rsrch_pp/ased_abt.htm)
- [20] AILLA: [www.ailla.org/pc/mainindex.html](http://www.ailla.org/pc/mainindex.html)
- [21] LACITO: [lacito.vjf.cnrs.fr](http://lacito.vjf.cnrs.fr)