The DOBES Archive: Its Purpose and Implementation

Hennie Brugman, Stephen Levinson, Romuald Skiba, Peter Wittenburg

Max-Planck-Institute for Psycholinguistics
hennie.brugman@mpi.nl

1. Introduction

The program for the documentation of endangered languages funded by the VolkswagenFoundation [1] started in September 2000 with a one-year pilot phase that included 8 documentation teams and one archiving team. This pilot phase was successfully turned over to the main phase in April 2002. The main phase now covers 12 documentation teams that have to deliver the documentation material to the archivist that is the Max-Planck-Institute for Psycholinguistics. It is not the purpose of this paper to describe the work of all teams within the pilot phase, the recommendations developed and the experiences made so far. For this we refer to the paper from Mosel, Dwyer and Wittenburg [2].

This paper is used to make statements about the views of the archivist within DOBES (DOkumentation BEdrohter Sprachen) with respect to the questions of the workshop organizers described in an online note [3]. To a large extent the statements being made in this paper were discussed at the three DOBES internal workshops. But the purpose of these statements is to raise an open discussion. Therefore, we cannot exclude the possibility that the linguistic teams do not share some of the views presented here.

2. Principal Function

The fundamental reason for the DOBES program is the fact that many of the currently about 7000 languages are endangered with many predicted to be extinct within a few years. This creates two tasks that pose different and partly conflicting requirements:

1. Members of the research community have a scientific and ethical obligation to use their skills to document some of these languages to preserve this cultural heritage for future generations.
2. The researcher community has an ethical obligation to aid indigenous attempts to revitalize the languages.

While the first goal is directed to long-term perspectives, the second goal has short-term aspects. Before making other statements this potential conflict has to be clarified. Let us briefly mention the parties involved[1]. We can speak about 5 main parties that have a role: (1) The funding agency has direct contacts to the documentation teams and the archivist. The primary focus of the program is documentation and not revitalization. (2) The members of the indigenous communities whose language performance is recorded and documented. They can act as users in the case of revitalization, since at that moment they will use the material gathered during the run of the project. Parts of the communities have expressed their wish to receive material back that they can use in education and cultural activities.

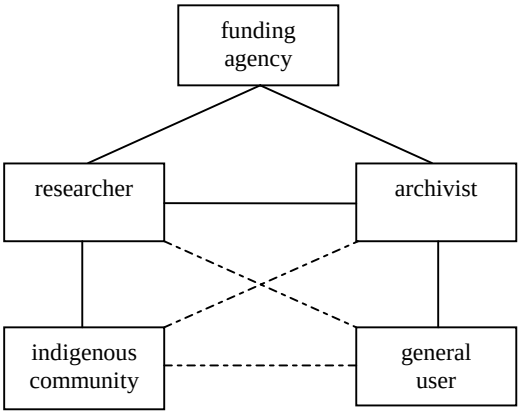

Figure 1 indicates the major parties involved in the DOBES program. Some have direct contacts due to the definition of the program, but there may be occasional direct or indirect contacts between other parties (indicated by stippled lines).

(3) The researchers who document the languages have contacts with the community members and who are faced with both fundamental goals. On the one hand they have to provide the documentation and on the other they have to help their community with revitalization. Third, being researchers themselves they are interested in creating publications, i.e. documentation itself is currently not accepted as serious scientific work for career advancement. (4) The archivist has to house the material, organize it following clear principles and guarantee persistence and accessibility. The archivist is in general not a linguist and not involved in the affairs of the individual communities. (5) There will be other users than those described beforehand who want to get access to the stored material.

Given this scenario the primary goal of the archivist is to store the incoming material in a safe way such that it can be accessed even after many years. It is not the task of the archivist to present the stored data such that it can be used for revitalization purposes. However, the archivist

[1] A more detailed analysis was made within the DOBES program to serve as basis for its attitude with respect to legal and ethical aspects. The documents will soon be available on the DOBES web-site.

has to help the researcher in this goal with technological procedures and advice. The archivist may decide to collaborate with linguists to present selections of the material to the public in order to make them aware of the lost or endangered cultural heritage.

So the principal role of the archivist is dedicated to the first of the two main goals and only secondarily to the second one. This may be seen differently for example by AIATSIS [4]. But as far as we can see, AIATSIS is not just defined as an archive, but has a much broader scope of activities with respect to the indigenous communities in Australia.

For the DOBES archivist it is clear that the archived material has to be available online. Otherwise, other parties cannot be supported efficiently with what they are doing. It has to adhere to the principles of modern information technology.

3. Pillars of Survival

If the archivist's primary task is focused on the persistence of the archive, then we have to define what the key pillars of survival are.

Generally speaking the survival is a matter of societal acceptance of the available material. This is given by

- political attitude with respect to cultural heritage;
- attractiveness of the content for the public after years;
- involvement of recognized institutions;
- cost efficiency of operation since societies normally have limited budgets for maintaining archives documenting the past;
- state of maintenance and degree of accessibility, interpretability and expandability of the archived documents;

When we put large human and financial efforts into the documentation of languages at this moment, we have the task to optimize on the conditions that influence survival.

The first two points can hardly be influenced. However, we know that societies develop differently, i.e. an optimization of the conditions means to spread the material across different areas around the world. With the documents we house there is no principal problem except that we have to consider the costs involved.

Important is the availability of recognized archiving institutions. For the area of Digital Libraries (multimedia language archives are DL) new structures have to be worked out. There is common agreement that traditional libraries and museums are not ready to house digital libraries with all its consequences. Therefore, we need intermediate solutions, since the documentation work has to be started now.

The DOBES archivist realizes that his task is a temporary one until new institutions have emerged which can take over the archive and offer a long-term perspective.

Cost efficiency is very difficult to achieve given the transient nature of today's technology. There is no way other than to keep the data online, since it is dynamic (people want to add annotations and comments of all sorts) and since people want to access it via modern technology anyhow. For every electronic archive one has to create several copies and to continuously migrate to new storage media[2]. On the other hand technologists have to look for a new type of medium that could be used to capture the static part of the data for longer time periods. Currently, however, no technology can be seen which does this. Under these circumstances cost efficiency can only be achieved by automatic procedures in larger computer centers.

For the DOBES program the archivist has declared that he will himself store three copies of all material with one copy not in his building. Further, all data will be mirrored at the MPI for Evolutionary Anthropology in Leipzig. Further, the MPI for Psycholinguistic will migrate the data to new storage media during the coming 6 to 10 years[3] to keep it available. Longer reaching statements cannot be given due to the scientific nature of the host institute. Procedures have been defined that can keep the archive as long as possible on a cost efficient and effective basis.

The last point mentioned in the list above has to with organizational and technical aspects that are dealt with in the next chapter.

4. Accessibility of the DOBES Material

4.1. Nature of Data

First, we have to describe the data we expect to get. Although parts of the data will not change, the data is dynamic. The DOBES pilot phase has made clear how gigantic the effort would be to create an exhaustive description within the limited period of time. Therefore, we can expect in general shallow descriptions with in depth analysis and description for some selected material. Since all data is online available, in particular the raw data such as audio and video recordings, we can expect that more researchers will elaborate on this material.

The data is dynamic and the archivist has to set up structures that allow easy updating and version control.

The data types we are confronted with will vary largely. Basically all sorts of material that was used in studying language and cultures will be generated and integrated. Media data will include video, audio, data from special phonetic recording devices, photos and others. Textual material will include annotations and lexica as major data types, but many different types of notes such as field notes, ethnographical notes and sketch grammars. The DOBES pilot phase showed that all data types are delivered in a large variety. Especially existing material comes along in out-dated formats. Also the preferences of the researchers for standard software such as MS WORD or EXCEL lead to documents that are in proprietary formats.

In DOBES it was agreed to convert all these data to a minimal set of standard formats that are up to date and have a high degree of openness.

---

[2] Currently, storage media are obsolete after 4 to 6 years, i.e. after a relatively short period of time the data has to be transferred to a new medium.

[3] Probably the material will be housed longer, but it is not clear which level of support can be given to access the data from the outside.

4.2. Standards in DOBES

After a exhaustive discussion between all teams the archivist choose to adhere to the following standards:

- audio:
  - 44.1 or 48 kHz linear coding, wav format
  - in future perhaps compressed formats (??)
- video:
  - until now MPEG1 compressed format
  - from now on MPEG2 compressed format
  - all other formats may be generated from MPEG2 to serve certain needs
- photos
  - any uncompressed format such as TIFF is optimal
  - people mostly come with compressed formats such as JPEG
- annotations:
  - XML as basic syntax
  - EUDICO Annotation Format (with an open and documented XML Schema as basis)[4]
  - intermediate support for Shoebox and Transcriber format
- lexica:
  - Shoebox accepted as intermediate format
  - XML as basic syntax intended
  - no schema developed yet
- notes:
  - plain text or HTML
  - XML to be accepted - yet no schema definitions
- metadata:
  - according to the IMDI definitions
  - definitions are open in form of XML Schemas
- character encoding:
  - rely on UNICODE

These standards are well documented, but there is no assurance that any of them will survive for more than the next 5 years. So what can the archivist do to guarantee long accessibility and interpretability?

The archivist has to make sure that there is documentation available about the chosen encodings and refer to those centers that have documented the encoding standards such as for MPEG2 and XML. It is not the task of the archivist to collect all information about the encodings at his site.

Everyone who has the permissions to use the data can choose his own way of processing them, since their encoding and format is described.

The archivist has to address the question what happens when new standards are emerging. The old ones may need to be replaced otherwise there is a danger that the data cannot be interpreted easily anymore after some time. Here, from cost efficiency reasons the archivist cannot convert again and again all material. But it might be necessary that the archivist takes over the responsibility to store himself the necessary documentation about a certain standard. So, if MPEG2 would become obsolete and there is a chance that documentation about MPEG will diminish the archivist should make clear that he has the necessary documents together with the data. The documents still would give everyone the possibility to construct algorithms that can interpret the data stream.

[4] When the AIF format and access APIs are stable and have the necessary expressive power and turn out to become widely accepted, the DOBES archivist will also support that format.

4.3. Resource Management

Resource management is an important task when maintaining a dynamic corpus of increasing size (currently about 250 GB of data). The DOBES teams were an active and important part in the discussions towards the IMDI metadata set. Using the IMDI metadata approach it is easily possible to integrate new data, to discover individual or groups of resources and to carry out operations on them.

Resource Management in DOBES is done on the conceptual level, i.e. IMDI types of metadata descriptions are used to organize the complete corpus. Therefore, a clear organizational principle is applied within DOBES.

4.4. Methods of Access

All data is kept online, i.e. if the user has the right permissions he can directly access all material via the networks. DOBES offers two ways to access the data (that holds for metadata as well as for the resources themselves): (1) The user can use the XML files, pars them and interpret the data. (2) The user can make use of shells that the DOBES archive offers. So the DOBES archive does not force the user to use certain tools.

For metadata the DOBES archive offers the IMDI BCBrowser as a simple shell allowing the browsing and searching in the linked metadata domain. A disadvantage for the general user is that he has to first download that tool to easily operate on the metadata descriptions. The archivist expects that future general-purpose browsers will give support for XML Schema definitions.

For annotated media files the DOBES archivist offers the user to download components of the EUDICO tool set to easily operate and search on annotated media files. But again: if the user believes that his tools are better he can decide to play for example the MPEG movies with whatever he has.

5. Other points of Relevance

Openness of Data

All metadata of the DOBES archive are open so that everyone can inform himself of what the archive is holding. With respect to the resources the access is primarily dependent on the attitude and interests of the indigenous community. There is a trend in the researcher community that all texts created by the linguists should become open - at least after a period of 3 years.

Analog Data

The DOBES archive does not want to house original tapes whether the recording technique is analog or digital. The researcher always gets the tape back and in addition a copy of the digital media file. The DOBES archivist only wants to take care of the digital versions on computers.

Misuse of Data

The DOBES archivist will treat the housed data very carefully, i.e. it will handle access permissions, logo insertion and watermarking with great care. However, there is no guarantee against misusage. The online availability increases the probability of misusage.

Insertion of a visible logo was tested, but the teams want their individual design. Watermarking has to be tested.

Interoperability

Interoperability is an issue especially as far as metadata is concerned. Search engines should operate on all holdings to give the user a quick and broad overview about the available data. The trend towards well-documented XML formats makes it possible to include XSLT scripts that do some online transformation on the fly.

6. References

[1] DOBES: http://www.mpi.nl/DOBES

[2] U. Mosel, A. Dwyer, P. Wittenburg (2002) Methods of Language Documentation in the DOBES Project. In Proceedings of the LREC 2002 Conference. LasPalmas Spain.

[3] Role of Archives: http://www.mpi.nl/lrec/

[4] AIATSIS: http://www.aiatsis.gov.au