Report of the Archive Advisory Board Meeting

17.11.2005

Peter Wittenburg
Daan Broeder

Participants: Peter Doorn (DANS), Jost Gippert (U Frankfurt), Laurent Romary (CNRS), Bernhard Neumair (GWDG), Dietrich Schüller (Phonogrammarchiv Wien), Harald Suckfüll (MPG-GV), Vera Szöllösi (VWS – Guest), Stephen Levinson (MPI), Peter Wittenburg (MPI)

Agenda

For the Archive Advisory Board (AAB) meeting the following agenda was accepted:

1.	Welcome	Stephen Levinson, Vera Szöllösi
2.	Tasks of the AAB	Peter Wittenburg
3.	Report to the AAB	Peter Wittenburg
4.	Discussion of the Report	AAB
5.	AAB internal matters	AAB

The discussion was based on two documents that were circulated in advance by Wittenburg:

- a proposed set of tasks of the AAB (Appendix A)
- a report about the current state of the archive and the technologies used (Appendix B)

Appendix C contains a short summary of the results of the DELAMAN meeting held in Austin at 21/22. November, since they may be of interest for the AAB members and add to some points.

1. Welcome

Levinson welcomed the AAB on behalf of the board of directors of the MPI. He gave a short description about the great endangerment of many languages and explained the obligation that also researchers feel to preserve the cultural treasure that is encoded in languages. He expressed the clear will of the directorate to maintain the language resource archive at the MPI in the future and explained the support for long-term archiving by the president of the MPS. The AAB is seen as a very important means to achieve a higher degree of stability of the archive and guarantee state-of-the-art procedures. So the board of directors is grateful to the AAB members that they are available to help the MPI.

Szöllösi welcomed the AAB on behalf of the VolkswagenFoundation (VWS). She explained the hope of the VWS that the AAB may help to guide in particular the DOBES Archive through the challenges of the dynamic technological development and to guarantee preservation and accessibility. She expressed the need to combine the knowledge of experts with different backgrounds as is the case for the AAB and finally thanked the AAB members for their efforts.

2. Tasks of the AAB

Wittenburg briefly summarized the document about the proposed set of tasks for the AAB. He also explained that the MPI houses several sub-archives and that they are all treated the same at the physical level. The distinctions are made at (1) the logical level (organization of the data with metadata), (2) at the level of archiving agreements (standards to be used etc) and (3) at the level of access management. So the AAB advice will be checked on its relevance for the whole archive.

The AAB briefly discussed the proposed tasks and agreed with them. In particular, the AAB should help to ensure that

- the archiving principles represent the latest technological state-of-the-art
- appropriate encoding and format standards have been chosen
- all decisions about major changes to the archive are to be made explicit
- the needs of long-term preservation and short-term access should be balanced
- possible extensions should not influence the stability and quality of the archive
- appropriate data protection measures are to be taken

It was briefly pointed out that there is also the Linguistic Advisory Board that also fulfills an important role in the DOBES programme. It is complementary to the AAB, since it focuses on all aspects that emerge from especially linguistic research aspects.

3. Report to the AAB

Wittenburg summarized the essentials of the current state of language resources archives, the technologies applied and the further future developments. (slides are available at the web-site).

4. Discussion of the Report

This part summarizes the points made by AAB members and gives answers of the MPI experts where applicable (*in italics*).

1. It was suggested to do Web-Statistics to have an idea about the usages of the archived material and it was asked who asked for resource access until now.
   W argued that statistics can be very misleading and that yet there is not yet much content in for example the DOBES archive. This will change when the first 15 teams will finish during the coming months. Then the archive will have to log the access to resources anyhow. This feature will be build in at the beginning of 2006.
   For the DOBES archive until now mainly journalists wanted to access the material for their work. Some teams want to have complete copies of their sub-archives.
2. It was criticized that the usage of GoogleEarth as one of the entry points for the archive does not respond to the users needs and that not much time should be invested. Others argued that the geographic information display is a means to link information from different disciplines and to improve the knowledge about the data. It was questioned whether dependencies with respect to Google can occur which are seen as dangerous.
   W responded that the GoogleEarth showcase just cost 3 days of work, but has an enormous potential of increasing the interest in archives and that new research paradigms will be made possible. MPI will continue to improve this small application and once settled for the MPI all DOBES teams will be asked how they would like to present themselves and their projects. GoogleEarth is just one of the possible platforms for offering selected language material. The way it is done is simple and allows a clear split between geographic and language information. It is agreed that dependencies on external information providers (Google) should not occur, with the current solution there is no reason for such fears since the language information is in XML and can easily be transformed to serve other geographic systems.
3. It was argued that as much data as possible should be made available to the interested community and that we need methods to be able to refer to archive material in publications.
   W responded that the decision is primarily with the responsible project leaders, since only they know the concerns of the language community members. It has to be seen how much of the data will become openly available finally. The issue of setting up a standard for referencing is seen as very important. The introduction of stable URID (Unique resource ID) can be seen as a step into this direction.
4. It was asked how the handling of raw material compares to handling annotations. It was also argued that in particular the language communities are primarily interested in accessing the primary material.
   This question is difficult to answer and depends on the direction of the question. For the researchers the creation of high quality annotations is a very time consuming process taking much more time that making the recordings. For the archivist much depends on the formats and tools used. Any "strange" format will require special treatment and can become very costly. Even the conversion from Transcriber producing XML to an XML format that is compliant with more generic XML annotation formats is difficult since it requires an interpretation by the archivist. Even more problematic is that for the more than 400 Shoebox files in the archive no typ and lang files are stored, i.e. no information about structure and

character encoding is stored. A conversion could be very costly. After the first 15 teams will have finished soon a better overview can be given.

5. It was asked what will be the state of validation of annotations and whether they are part of the primary resources in an archive. Others argued that the original transcriptions and annotations indeed are an integral part of the documentation, since the raw material would otherwise not be interpretable.
   It was agreed that original annotations are almost equally important than the recordings and that they therefore have to be part of the archive. Of course, they may change over time, but this should not mean that new versions may affect the existing data. The quality is primarily left to the responsible researcher.
6. Is data enrichment or modification a goal of a digital archive and if so how to ensure the persistence of the original data.
   Data modification is a goal for several reasons: a) the researcher involved does not know everything and will make "errors". It is desirable that enrichments can complete unfinished analysis; b) enrichments are chances to include the language community and young students; c) a documentation will not lead to a complete analysis therefore other researchers can add information to make the description more complete. Of course, it has to be taken care that enrichments are kept separate from the original data.
7. It was asked whether it is better to store DV instead of MPEG2 as back-end video format. It was added that perhaps re-digitization will be necessary, in this case it is necessary to store the original tapes.
   This issue will have to be analyzed again. MPEG2 was chosen, since it is an open format used by many and since it offers a comparable quality at a much lower capacity requirements. DV is a proprietary format and can be changed at any time. But the quality issue will be discussed between the Phonogrammarchive and MPI.
   Indeed re-digitization already occurred after the first year where the DOBES teams to decided to step over from MPEG1 to MPEG2. The MPI has no special facilities to store tapes for a long time and guarantee readability. But it was agreed that this question has to be addressed as long as we don not store uncompressed video.
8. The list of accepted formats should not be fixed, but should be open to new formats.
   This was agreed and the list of accepted formats was already changed. On the other side one of the most important principles of archives should be that changes should not occur too frequently.
9. It was questioned whether it would make sense to separate the DOBES archive from the MPI archive in all respects and what the respective costs would be? It was also questioned how deposits and conversions to archival formats from other people should be funded? Some AAB members argued that a physical separation cannot be the goal and that archives such as at MPI have to be open to other depositors. The need of such open strategies is clearly supported by the fact that about 80% of the important data is endangered. This data is typically not contained in archives and kept in inadequate circumstances.
   Some DOBES members have the impression that not enough money is spent on servicing the DOBES teams. W stated again that this is not true given all the services the MPI delivers. One of the main DOBES contact persons for example, Paul Trilsbeek, is funded 100% by the MPI and has a permanent job. This is necessary since the required stability can only be assured by permanent staff. Therefore, technically and personally it does not make sense to separate the sub-archives – it would create more overhead. Also the maintenance of lists that allow a detailed calculation of work times does not make sense – how should MPI calculate the private time of team members for example. Any separation would fail und would not be fruitful.
   For the reasons mentioned above, the MPI will open its archive. Simple deposits from other projects do not cost big money, since they have to be done via the LAMUS ingest software and all storage space, server capacity etc is funded by the MPI. We will also carry out digitizing tasks for others and will take care that this is not paid by DOBES money. However, we should also say there are no clear boundaries for all requests. Members from the Seifart project, for example, brought a whole box of highly valuable recordings that had to be processed in shortest time possible, due to transportation problems. (Therefore, this was put

high on the priority list due to the timing constraints and we did not look who is doing what in this case). For these kind of external contributions we will also not require standard formats etc, i.e. if people come and deliver WORD files for example, we simply will put them into the archive. In some cases we may run existing programs to convert the material, but here we will also take care that this will be done by MPI people. The person doing most of the conversion scripts is also permanent staff member of the MPI.
Due to these different requirements, however, we have to follow a procedure for depositors who want to add data to the DOBES sub-archive (we already had one request from a training workshop participant). Such a request has to be discussed by the DOBES Steering Board and after the end of the programme we will ask the linguistic advisory board.
Summarizing, we can say that the MPI will continue with its open deposit policy and that W has to take care that this will not lead to a decrease in service quality for the DOBES projects.

10. It was argued that storage costs will become marginal due to technology change. After every innovation cycle the current data volumes will only take a small fraction of the then available capacities.
    This statement was widely accepted, however, generating and storing of uncompressed video is still not tractable given the current constraints. This may change dramatically when holographic technology will become available in a few years time.
11. It was questioned whether it can be expected that researchers will continue to put efforts in the classification of their data, i.e. create metadata.
    W sees no other chance than to motivate in particular young researchers to create the necessary metadata, since they are the glue for all resources and the basis for any discovery. Only discipline at this side will prevent a completely useless and inaccessible archive. MPI will continue with its strict policy.
12. The MPI archive should demonstrate that it is OAIS compliant.
    This as agreed and a document will be written and published at the web-site.

5. AAB Internal Matters

A number of points were agreed and they are briefly listed.

- Laurent Romary (CNRS) will act as the chairperson of the AAB.
- The AAB documents will be open.
- The MPI will write a first version report which will be exchanged and corrected. The final version will be put on the web.
- Together with the chairman an appropriate text will be created for the web-site.
- The AAB will be consulted about essential changes in the setup of the archive. The chairman will be the contact person to give first advice.
- The AAB will meet in person or virtually (via video conferencing) once per year.