A Grid of Regional Language Archives

Paul Trilsbeek, Daan Broeder, Tobias van Valkenhoef, Peter Wittenburg

Max Planck Institute for Psycholinguistics

Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

E-mail: {Paul.Trilsbeek, Daan.Broeder, Tobias.vanValkenhoef, Peter.Wittenburg}@mpi.nl

Abstract

About two years ago, the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, started an initiative to install regional language archives in various places around the world, particularly in places where a large number of endangered languages exist and are being documented. These digital archives make use of the LAT archiving framework [1] that the MPI has developed over the past nine years. This framework consists of a number of web-based tools for depositing, organizing and utilizing linguistic resources in a digital archive. The regional archives are in principle autonomous archives, but they can decide to share metadata descriptions and language resources with the MPI archive in Nijmegen and become part of a grid of linked LAT archives. By doing so, they will also take advantage of the long-term preservation strategy of the MPI archive. This paper describes the reasoning behind this initiative and how in practice such an archive is set up.

1. Introduction

Currently about 6500 languages are spoken worldwide. Even optimistic views assume that at the end of this century at least 50% of these languages will be extinct in the sense that there will be no more active speakers of the language left. Given this trend, caused by many different factors but in particular by the globalization of economy and information distribution, several initiatives such as DOBES [2], HRELP [3], PARADISEC [4] and ELF [5] have been launched during the last decade to document these languages that are highly endangered before it is too late. The documentation serves essentially two goals: (1) To document the rich linguistic variety and an essential part of our cultural heritage for future generations and (2) To make use of the documented material for setting up revitalization programs where possible in order to maintain linguistic and cultural diversity. For almost all these documentation programs, an archiving strategy has been defined which includes the establishment of a central archive where copies of the recorded material have to be deposited.

A challenge that all digital archives are faced with is the long-term preservation of the digital data due to the rapid changes in storage technology and the relatively short lifetime of digital file formats and encodings. In the case of the DOBES archive, which is based at the MPI for Psycholinguistics, a long-term preservation strategy has been worked out that consists of a number of policies. First of all there is a policy to migrate the technical server and storage infrastructure at the MPI to the latest state of the art every 4 to 5 years, before it becomes obsolete. The second policy is to distribute the data to a large number of physical locations, while taking property and access right issues into account. There is a data backup agreement with two large data centers in Germany and one sister Max Planck Institute, which means that there are a minimum of 7 copies for each file in the DOBES archive in 4 different geographical locations in Germany and the Netherlands. Efficient file synchronization protocols are used to create the various backup copies and to update them dynamically. The third component has to do with the use of file formats and encodings. The DOBES archive limits the number of file formats and encodings that can be used for audio, video, images and various written resources and uses open formats as much as possible. This will make it more feasible in the future to convert data to a different format or encoding in the event that a format has become obsolete. It may be apparent that such a long-term preservation strategy requires large investments in network, server and storage technology at regular intervals. Establishing a digital archive with a long-term perspective therefore can be problematic particularly in those countries where we still find large linguistic diversity and where we urgently need to document endangered languages such as in South America, in Africa and in Southeast Asia. The result of this is that - like in the old colonial times - cultural heritage is transferred to western countries and stored far away from places where it originated.

To compensate for this deficit, the MPI for Psycholinguistics started a program to set up so called “regional” archives. In this program, servers with enough online storage capacity are installed at sites that are close to the regions where the recordings were made or where the local researchers are located. By exchanging data with the central archive in Nijmegen, the regional archives automatically become part of the long-term preservation strategy of the MPI.

2. Setup of the Archive

The procedure for setting up such regional archives and getting them into the operational phase includes a number of steps: (1) A discussion needs to take place about the organizational embedding, the responsibilities, the local technological infrastructure and about the agreements to be signed. (2) A visit is scheduled to setup the server and to integrate it in the local IT infrastructure. Two scenarios can be thought of - a remote installation of the archiving software by an expert in Nijmegen is feasible if there are system administration experts available locally. Otherwise an expert from Nijmegen needs visit the site and do the work there. (3) The data is copied, the catalogue is put online and an entry is made in the Google Earth overlay. (4) A training course is given for the local system administrators and archive managers and optionally another course for users of the archive. (5) A data copying policy is decided upon and configured so that the local archive managers control what data will be synchronized with the MPI or other LAMUS managed archives.

Generally when a request for setting up a regional archive comes in, the first thing that needs to be decided upon is an appropriate location for the archive. Even though the technical demands are not that high, there needs to be some basic technical infrastructure in place, such as a stable supply of electricity, a broadband internet connection, and basic technical support to keep the server running. Additionally, in hot and humid climates, a climatized room is needed. The servers that are being installed to form the digital archives are equipped with the LAT framework [1] of creation, archive management and web-based utilization software that has been developed at the MPI and therefore offer the same functionality as the archive at MPI. An important part of this framework is the IMDI metadata scheme that has been developed specifically for linguistic resources (Broeder et al. 2001, [6]) Generally the relevant linguistic data for the particular region is being copied to these servers, obviously in agreement with the corresponding documentation teams. The servers are under full control of the local experts, however, a formal agreement is needed that handles the exchange of data and the access to the servers by the MPI experts for debugging and software maintenance purposes. For the DOBES collections, an agreement is useful to ensure that the MPI stores an up-to-date copy of the material, as this is a requirement of the DOBES program. All regional archives are added to the MPI's "Language-Sites" overlay [7] for Google Earth, by means of which users can reach the catalog of the regional archive by clicking on a link in the Google Earth interface (see van Uytvanck et al. 2008). The regional archive's catalogue can also be made available directly from the MPI's catalogue, in which case it will also be part of all catalogs that harvest the IMDI metadata.

From what we have experienced with the archives installed so far, the installation of the LAT software can be done within a couple of days if the specifications for the machine and the operating system (Suse Linux Enterprise Server) are met. The training of the local experts in how to administer the archive generally takes at least a few days, assuming that basic UNIX knowledge is already present. So generally our specialists are around for 5 days to do the installation and to train system administrators and archivists in how the tools work and how to manage the archive.

The concept of synchronized regional archives is based on the LAMUS (Broeder et al. 2006, [8]) archive management system and the aforementioned IMDI metadata infrastructure, both components turn out to be increasingly stable and robust in the differing circumstances. Currently, we are working on a LAMUS component that allows us to dynamically synchronize (sub)-corpora between the archives at the logical level and not at the physical level as would be the case when using backup or file synchronization software. During the first half of 2008 we will make a

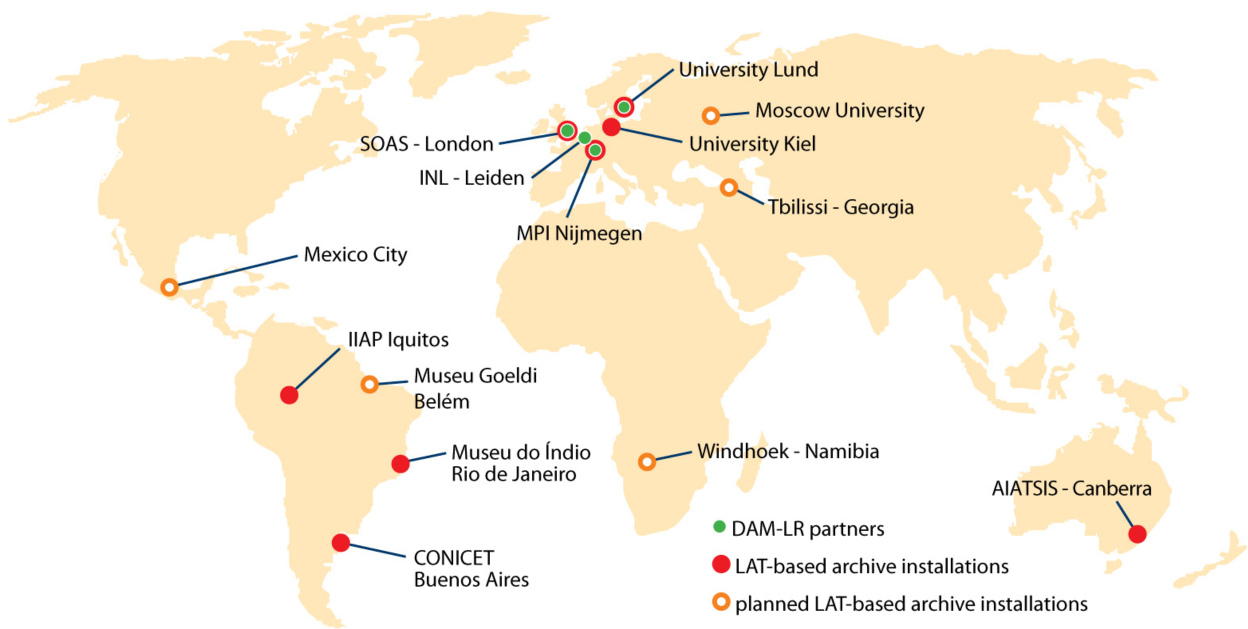

Figure 1: Distribution of the current LAT-based archives and planned archives for 2008.

synchronization component available that enables cross-archive synchronization on the basis of the metadata information and that takes important features into account such as maintaining versions, maintaining unique and persistent resource identifiers and maintaining the same access rights at both sides without the need for human intervention. Currently, we only know of one other repository/archive system that offer these features, the “Storage Request Broker” [9], which is an excellent framework mainly used in physics, but which lacks the kind of special support for our application domain. LAMUS on the other hand is a system that could in principle be used for a large variety of resource types; its power however is given by the specialized functions it offers for analyzing and visualizing the stored language resources.

The researchers at the regional centers can use the new server for any kind of data that fits into the LAMUS/IMDI scheme, which actually can incorporate a broad variety of resources. By default, the LAMUS system accepts the data formats that are accepted by the MPI archive, however this could be configured differently for the regional archive. One thing that one has to keep in mind though is that a large variety of data formats in an archive can reduce the chances of long-term survival of the data, because conversions to newer formats may be too costly. The agreements for data exchange between the regional archive and the MPI could be made such that all newly added data of the regional archive is copied to the MPI. This is important, since many countries do not have an infrastructure available to create several copies of the data and to take care of long-term preservation aspects.

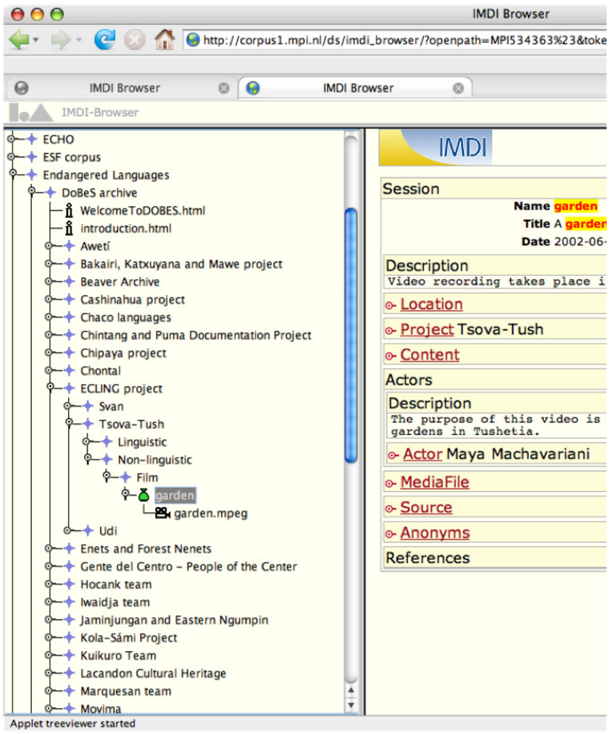

Figure 2: IMDI archive browser and metadata viewer gives online access to all archived resources.

3. Current regional archives

Currently, seven LAT-based archives have been established in various places around the world; two of them at partner institutes within the DAM-LR European project [10] and five regional archives. We have already received requests for setting up ten more, five of which we are planning to do in 2008 (see Figure 1). How the various regional archives are embedded in national structures is very different. In Argentina and Brazil for example it is the intention to make these archives into national centers for language resources of all kinds. In fact, in Brazil it is planned to have a national network of at least three such centers in the big cultural institutions dealing with indigenous cultures and languages. In most countries or regions these new language archives are starting to become centers for working with languages, for creating “Live Archives” [11] in which people add various types of enrichments to the stored collections. These enrichments can be new resources, new annotations, commentaries on existing material or “knowledge spaces” in which culturally relevant concepts are related to each other and have links to all sorts of resource fragments in the archive. In many cases, having the language material physically stored in a nearby location certainly makes a difference in the attitude that both the members of the speech community as well as the local linguists have towards archiving the material. The internet as an access and enrichment medium however still remains abstract for many when working with primary data, for example most researchers are still thinking about their work in terms of a "local" or "download first" working paradigm. This way of working can and should of course also be supported by the archiving framework, certainly for tasks that can not easily be done over the internet in a precise way and tasks that need to be done in the field.

From our experience with seven installations in the last two years it became clear that the LAT software machinery is robust and stable now. Nevertheless, a local archive manager is needed to maintain the consistency of the archive and to help out depositors of new material. On the LAT software side, continuous maintenance updates and functional extensions are taking place; therefore an automatic update mechanism needs to be developed such that all LAT archives can be updated to the latest versions of the software. Also an information exchange network for the various archivists involved needs to be set up in order keep them up to date about the latest developments.

A recent development that we started which is related to the regional archives initiative is to facilitate the creation of so-called “community portals”. The interfaces of the LAT software require a fair degree of familiarity with concepts of modern computer operating systems such as navigating a folder structure and making use of the right mouse button to show a contextual menu (see Figure 2). Also, the way the linguistic corpora are organized is often from a linguistic point of view. Therefore, the standard way

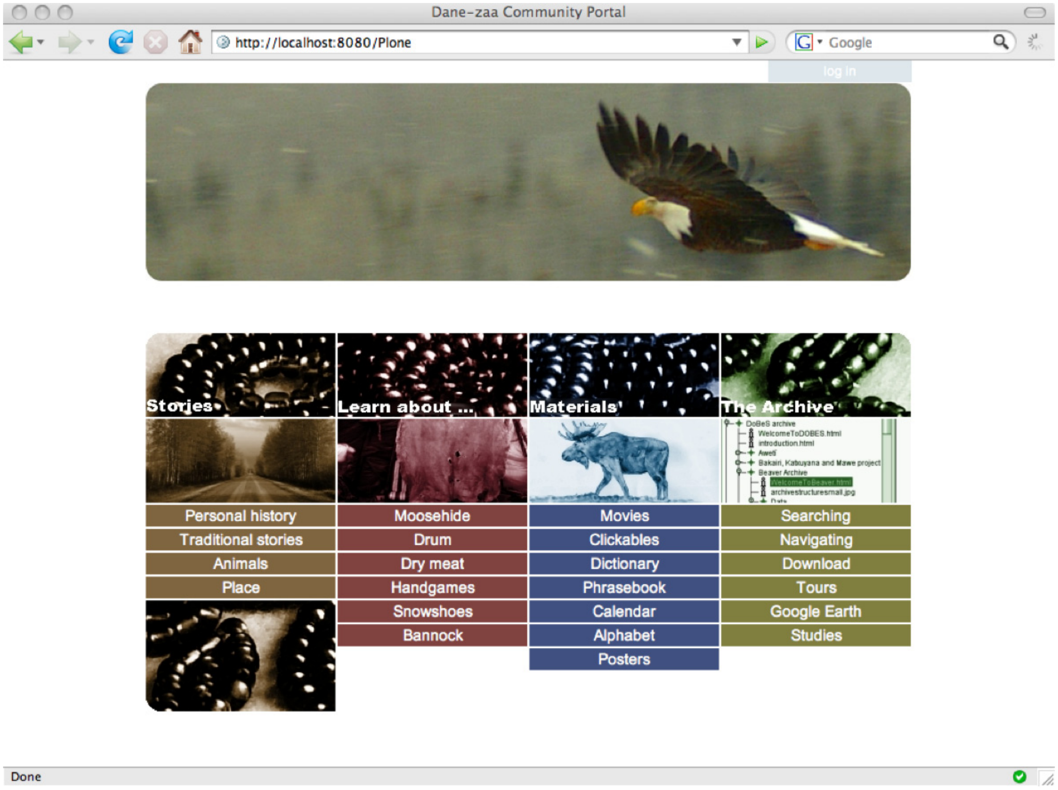

Figure 3: Prototype of the Beaver Community Portal.

in which the LAT tools give access to the data is often not that well suited for members of the language communities. To offer a more attractive interface for the language community for the discovery of archived material, a “community portal” is more suitable. Such a portal offers a nice graphical interface in which audio and video recordings can be easily browsed by categorizing them by genres that make sense for that specific community. At the MPI, a framework was developed that combines a standard web Content Management System (Plone) and a REST search interface for the IMDI metadata. Figure 3 shows a prototype of a Community Portal that was designed by the documentation project of the Beaver language spoken in British Columbia and Alberta in Canada.

4. Conclusion

The initiative to start the "regional archives" program turns out to be very important since it puts the data closer to where it was recorded, which increases the feeling of responsibility and involvement of the local researchers and the speech community members. It opens up the possibility of including many resources that are digitized in an archiving framework that is aimed at long-term preservation, something that would be rather difficult to achieve if a system needed to be developed from scratch and if no connections to an institution such as the MPI could be established.

5. References

Broeder, D., Offenga, F., Willems, D., & Wittenburg, P. (2001). The IMDI Metadata Set, Its Tools and Accessible Linguistic Databases. *Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia*, 11-13.

Broeder, D., Claus, A.,Offenga, F., Skiba, R., Trilsbeek, P., & Wittenburg, P. (2006). LAMUS – the Language Archive Management and Upload System. *Language Resources and Evaluation Conference 2006, Genoa.*

van Uytvanck, D., Dukers, A., Ringersma, J., & Trilsbeek, P. (2008). Language-Sites: Accessing and Presenting Language Resources via Geographic Information Systems. *Language Resources and Evaluation Conference 2008, Marrakech.*

[1] http://www.lat-mpi.eu
[2] http://www.mpi.nl/DOBES
[3] http://www.hrelp.org
[4] http://www.paradisec.org.au
[5] http://www.endangeredlanguagefund.org
[6] http://www.lat-mpi.eu/tools/imdi
[7] http://www.languagesites.org
[8] http://www.lat-mpi.eu/tools/lamus
[9] http://www.sdsc.edu/srb/index.php
[10] http://www.mpi.nl/dam-lr
[11] http://www.mpi.nl/dam-lr/lra-flyer/lra.html