Large Multimedia Archive for World Languages

Peter Wittenburg
MPI for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen
peter.wittenburg@mpi.nl

Paul Trilsbeek
MPI for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen
paul.trilsbeek@mpi.nl

Przemek Lenkiewicz
MPI for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen
przemek.lenkiewicz@mpi.nl

ABSTRACT
In this paper, we describe the core pillars of a large archive of language material recorded worldwide partly about languages that are highly endangered. The bases for the documentation of these languages are audio/video recordings which are then annotated at several linguistic layers. The digital age completely changed the requirements of long-term preservation and it is discussed how the archive met these new challenges. An extensive solution for data replication has been worked out to guarantee bit-stream preservation. Due to an immediate conversion of the incoming data to standards-based formats and checks at upload time lifecycle management of all 50 Terabyte of data is widely simplified. A suitable metadata framework not only allowing users to describe and discover resources, but also allowing them to organize their resources is enabling the management of this amount of resources very efficiently. Finally, it is the Language Archiving Technology software suite which allows users to create, manipulate, access and enrich all archived resources given that they have access permissions.

Categories and Subject Descriptors
J.5 **[Arts and Humanities];** Linguistics

General Terms
Management, Documentation, Design, Reliability, Security, Standardization.

Keywords
Multimedia Archive, Lifecycle Management, Metadata, Long-term Preservation, Data Replication, Standards, Multimedia Access.

1. INTRODUCTION
In history languages and cultures were always changing due to many well-understood factors such as political and economic factors. However, in recent decades the dynamics of change got an enormous speed up due to globalization with the consequence that currently about one language is becoming extinct every week and that even major languages are changing. As in biology we see a huge decrease in linguistic diversity [1]. Not only languages are dying, but cultures are changing rapidly, identity building for young people becomes very difficult and the stability of societies is affected. We are deemed to loosing part of our cultural heritage since every language can be seen as a unique result of evolution resulting in rather different language systems. We also risk losing much of our knowledge about environment, species etc. since this is to a large extent being encoded in the semantics of a given language.

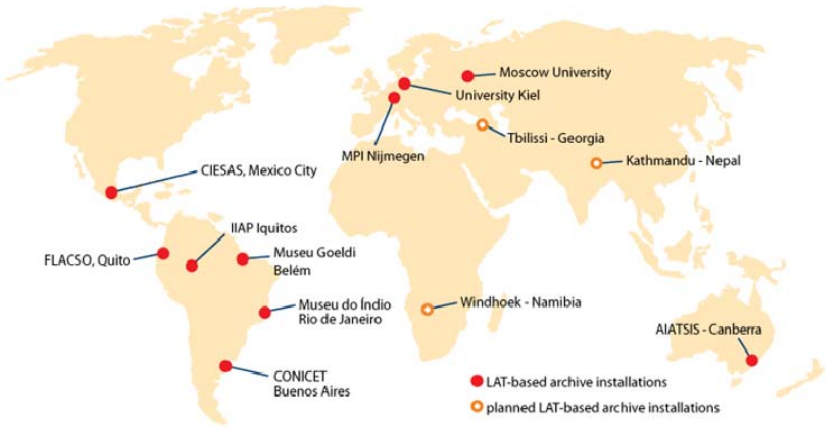
During the last decades we recognize an increasing awareness about these threats resulting in a number of world-wide initiatives to document, archive and revitalize languages (DOBES, HRELP, PARADISEC, ELF, AILLA, [2,3,4,5,6] etc). Of course these three tasks depend on each other since revitalization programs will depend on archived material which has been recorded by field workers beforehand. Also it is well-understand now, that we have the obligation to preserve our material and knowledge about languages for future generations, since they may want to understand their roots. Also we may wonder that like with seed banks that are created to allow future people to return back to "original" species, future generations may want to return to proper linguistic constructions that are currently blurring or which we currently are losing.

While a couple of decades ago most of the language documentation was driven by linguistic curiosity to understand another new linguistic system with all its special constructions, this has now changed completely. Linguists understood that documenting a language needs to be based on audio and video recordings that also show the cultural background on which languages were spoken and that some of these recordings need to be translated and annotated at linguistic level. Additional material such as lexica, sketch grammars etc. can help to give a comprehensive impression of the language.

During the last decade also the awareness has grown that making recordings etc. alone is not sufficient to guarantee that future generations will indeed be able to access the data. A UNESCO survey [7] has shown that about 80% of our material about languages and cultures is highly endangered due the fact that the electro-magnetic/optical substrates of the carriers that are normally used are deteriorating. In addition we know that players (hardware, software) for old formats (tapes, texts, media etc.) are not supported anymore after relatively short periods of time requiring very expensive curation.

Based on these facts the DOBES program, starting in 2000, designed its strategies which were oriented to multimedia documentation, intentions for the usage of standards, and goals for long-term preservation and accessibility. In this paper we will focus on lifecycle management, archive organization and access issues.

SSCS'10, October 29, 2010, Firenze, Italy.



This figure indicates which regional repositories have been set up in collaboration with the local experts. All these repositories exchange sub-collections of the archived resources serving two purposes: (1) taking care of data replication and (2) bringing back the resources to the regions where they are recorded.

2. ARCHIVE

Already a decade ago it was obvious that we need a proper organization and description of our resources to serve three major purposes: manageability of large amounts of related resources, discovery of resources and usage of metadata for scientific analysis. The IMDI metadata framework [8] was built as a result of discussions between linguists and technologists that satisfied all criteria in so far as it

- makes use of an element set and vocabularies that emerged from linguistic considerations and semantics
- allows to build hierarchies and collections for management and virtual collection building purposes
- allows to browse in hierarchies and search on descriptions
- offers a gateway to DublinCore to allow OAI-PMH based harvesting

Later the LAMUS archive management tool [9] was built that makes use of IMDI for archive management, allows users to upload new resources or resource collections, set access permissions based on linguistic needs and carries out checks on metadata correctness, on the consistency of all links and on the adherence of resources to the set of accepted formats. In the mean time persistent identifiers (PID) automatically registered with a Handle System [10] server were added to make the references independent of all changes in the storage configuration, i.e. also when new resources or collections are uploaded every object will be associated with a PID and the PID itself is associated with an MD5 checksum information to allow authenticity checks. To strictly maintain archivable formats no encapsulation was accepted, i.e. all resources including the metadata descriptions are stored in standard formats in the file system. Only for fast access purposes databases and indexes are created and used. This makes access to resources and their interpretability completely independent of layered software which is important for long-term access.

A local storage system was built that stores two copies of all resources, i.e. at upload immediately two copies are being created. Core of the storage system is the SAM-FS [11] hierarchical storage management system which manages fast disk array caches for the small textual resources, indexes etc., slow disk array caches for the media files and a tape library based on LT04 technology. With two servers and a double path SAN configuration single points of failures have been avoided. The archive currently stores more than 50 Terabyte of data contained in about 1 million objects. Since these two local copies will not be sufficient to speak seriously about long-term preservation dynamic copies are being created to two large computer centers in Germany at distinct location each of them having an agreement with another big computer center about long-term archiving. Thus all archived objects are available in 6 copies and to decrease the probability of failures two different dynamic replication protocols at physical level are used: with one center rsync [12] and with the other Andrew File System [13] based exchange are being used. In addition it is important that the president of the Max Planck Society has given a 50 years institutional guarantee for all resources stored at the computer centers.

Since all components operated very smoothly and as a whole[1] error free for several years, we can claim that we indeed take care of long-term preservation – at least as good as it is possible these days. This claim was confirmed by getting the Data Seal of Approval [14] after a formal assessment of our procedures.

3. NETWORK OF ARCHIVES

The MPI archive is member of the international DELAMAN network [15] of archives in the area of endangered languages and music and in the centerNet initiative [16] to use these platforms to synchronize about standards, procedures and policies. One of the outcomes of these discussions is the conviction that the digital area fundamentally changed the way to do archiving. In the area of analogue media it was obvious that every operation (copying,

[1] A down-time of one of the centers not allowing dynamic replication for a while is not dramatic.

viewing) would decrease the quality of the material (carrier and/or content). Therefore the big film companies decided to put copies of their master films into an old mine and to not touch this material. For digital data the opposite is true. We need to touch the digital copies regularly to carry out all sorts of tests and migrations (carrier, formats) to ensure interpretability. This can only be done since we understood that copying digital content does not decrease its quality. It was also understood that there has to be now distinction anymore between an interactive repository and an archive which is basically not accessed.

In addition the MPI team has setup 10 regional repositories all over the world as is indicated by the following diagram. The reason for doing this is twofold: (1) It is important to return the data into the areas where it was recorded to give the local people the chance to use and enrich it. Also the physical existence of a server with the data at a local institution gives a completely different attitude to data as if it would only be available virtually via the Internet. (2) Copying sub-collections of the data and spreading them worldwide will increase their chances of survival.

4. STANDARDS

Very important for long-term interpretability is the adherence to standards where possible. Here the MPI team participated in particular in the ISO TC37/SC4 [17] committee to work on the following issues: (1) ISO 12620 [18] as a model for registering data categories (formal concepts) and building the ISOcat [19] software to host the definitions many of which have already being entered; (2) Lexical Markup Framework [20] to have a generic model to represent all kinds of digital lexicons; (3) Establishing principles for associating persistent identifiers with linguistic resources; (4) Defining a set of generic guidelines for annotation formats. Of course widely accepted vocabularies such as the ISO language codes ISO 639-3 [21] are supported.

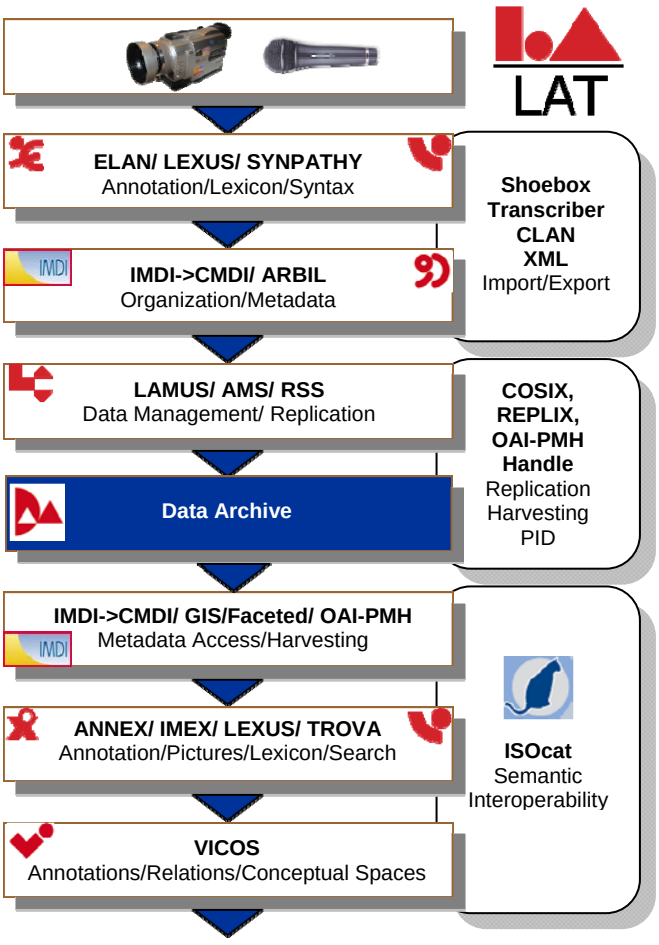
In addition we decided to adhere to a number of basic standards such as UNICODE and XML for texts, MPEGx for video representation, linear PCM with high quality for audio streams. The dynamics in the area of video codecs made it necessary to change our strategy 3 times in the last decade. When we started just MPEG1 were usable. Then we turned to MPEG2 as archiving format and are using increasingly often MPEG4/H.264 as presentation format. Recently after deep investigation we have chosen to smoothly turn over to lossless mJPEG2000 to finally have a master format from which we can create other formats without risking concatenation effects.

Since it is known that curation costs grow over time we apply an immediate conversion policy where possible. Since many tools still do not support standards and are not restrictive with respect to structures the conversion of for example complex lexicons is fairly cost intensive and not feasible without manual intervention.

5. ACCESS TO ARCHIVED MATERIAL

With the open source LAT (Language Archiving Technology) software [22] suite we have been developing software components that cover the whole lifecycle of language resource of different types without claiming that (a) these need to be used and (b) they include all functionalities. There are tools to create multimodal annotations for media recordings which can include time series such as eye tracking data, eeg data, etc. as well, complex lexica allowing to include multimedia fragments and syntax trees. The IMDI components allow users to create metadata descriptions that adhere to the IMDI schema and the associated vocabularies. The old IMDI components are currently being replaced by modern tools such as ARBIL [23] that combine metadata creation with organization capabilities and thus increase metadata creation efficiency. In addition we stepped away from a fixed schema approach, but let people now create their own profiles as long as they are using elements registered in ISOcat – the certified concept registry – which is of course important for semantic interoperability.

As already indicated LAMUS and its components for access management and access requesting are acting as gate keepers for the archive to ensure consistency and coherence, to associate PIDs, to create presentation formats such as MPEG4 for video streaming, to update fast search indexes, etc. A first component called COSIX has been integrated to do data replication and synchronization based on logical level (in contrast to the replication at physical level used for example by rsync) which allows us to properly exchange sub-collections with the regional



This figure shows the open source Language Archiving Technology software suite that has been developed to support the whole lifecycle and to help the researchers to create, manipulate, upload, manage, and enrich language resources. In collaboration with ISO the ISOcat software has been developed to allow users to register domain concepts and therefore facilitating interoperability.

archives. Together with the DEISA [24] project that brings together the high-performance computer centers in Europe we are working on the REPLIX [25] framework for safe data replication which is being based on policies at various levels.

A number of web-applications have been developed to be able to access the archived material via the web. The metadata is offered in various ways: (a) as the IMDI catalogue; (b) via IMDI search (simple and complex), (c) as an overlay in Google Earth [26] and (d) via a faceted browser in the Virtual Language World. Metadata selections can be used by these techniques which then can be used to carry out a content search via the TROVA [27] search engine. As well annotated media streams and multimedia lexica can be viewed and manipulated to a certain extent. VICOS [28] allows users to create conceptual spaces by drawing relations between lexicalized concepts, to navigate in this semantic domain and to open related archived resources from every node.

Thus the LAT software offers a comprehensive set of generic access technologies to those users who have access permissions to the content. Of course metadata is open.

6. CONCLUSIONS

The MPI team is housing one of the largest multimedia archives containing material about languages and cultures which to a large extent cannot be created any longer due to the rapid changes they are undergoing and even due to their extinction. Thanks to clear strategies from the beginning, in particular the IMDI metadata concept, the archive is in a consistent and coherent state although we have more than 200 internal and external contributors working independently and partly at different locations worldwide. The LAT software supporting widely accepted standards offers components that can be used to create, manage, replicate and access the archived data. The archive is unique in so far that it has 6 full copies at 3 different locations and in addition copies of sub-collections at 12 remote repositories that are synchronizing their data with the central archive.

The participation in international networks and initiatives is very important to fine-tune the strategies, to maintain cutting-edge technology and to integrate at least the metadata in portals such as the Virtual Language Observatory [29] which is maintained by the European research infrastructure initiative CLARIN [30].

7. REFERENCES

[1] D. Crystal (2000). Language Death, Cambridge University Press.

[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.ailla.utexas.org/site/welcome.html

[7] D. Schüller (2004). Safeguarding the Documentary Heritage of Cultural and Linguistic Diversity. Language Archiving Newsletter. Vol. 1, Nr. 3

[8] http://www.mpi.nl/IMDI/

[9] http://www.lat-mpi.eu/tools/lamus/

[10] http://www.handle.net/

[11] http://www.init-bs.com/download/SunStorageDay_SAM-FS-Roadshow.pdf

[12] http://de.wikipedia.org/wiki/Rsync

[13] http://www.openafs.org/

[14] http://www.datasealofapproval.org/

[15] http://www.delaman.org/

[16] M.A. Windhouwer, S.E. Wright, M. Kemps-Snijders. *Referencing ISOcat data categories*. Accepted for a presentation at *LRT standards* workshop . Malta, May 18, 2010.

[17] http://www.tc37sc4.org/

[18] http://www.isocat.org/

[19] http://www.lexicalmarkupframework.org/

[20] http://www.sil.org/ISO639-3/

[21] http://www.lat-mpi.eu/tools/

[22] http://www.lat-mpi.eu/tools/arbil

[23] http://www.deisa.eu/

[24] http://www.mpi.nl/research/research-projects/language-archiving-technology/replix

[25] http://earth.google.com/

[26] http://www.lat-mpi.eu/tools/annex

[27] http://www.lat-mpi.eu/tools/vicos

[28] http://www.clarin.eu/vlw/observatory.php

[29] http://www.clarin.eu/