

# Language Archive Management and Upload System

Peter Wittenburg, Daan Broeder, Andreas Claus

Max Planck Institute for Psycholinguistics  
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands  
[peter.wittenburg@mpi.nl](mailto:peter.wittenburg@mpi.nl), [daan.broeder@mpi.nl](mailto:daan.broeder@mpi.nl), [andreas.claus@mpi.nl](mailto:andreas.claus@mpi.nl)

## Abstract

When a language archive reaches a certain size the need arises for an automated system to manage this archive. One way is to create a tool for the researchers which allows them to manage the content of the archive themselves. Such a system should guide the researchers and allow them to insert new data and delete or change existing data. At the Max Planck Institute for Psycholinguistics we are developing software which fulfills these requirements on the basis of the IMDI Metadata framework. The requirements and possibilities of such a system are discussed in this document.

## 1. Introduction

At the MPI for Psycholinguistics a digital language archive has been set up over the last years that now covers approximately 40.000 sessions or resource bundles and other linguistic data types such as lexica, field notes and sketch grammars. Usually resource bundles combine different media files such as video recordings, audio recordings and in some cases other signals such as eye movement recordings and several layers of (multimodal) annotations. Therefore, the language archive already contains far more than 100.000 individual language resources. The contributions to this archive come from different projects all of which have very different objectives, such as:

- the DOBES programme, founded by the Volkswagen Foundation where 24 documentation teams work all over the world to document languages that will become extinct in a few years time;
- the corpus of the ESF funded programme on adult language learners
- the Dutch National Spoken Corpus that covers annotated spoken utterances with about 10 Million words
- the corpora from the language and cognition department of the MPI that covers various gesture studies, interviews, recordings of elicited utterances and others in many different languages
- the corpora from the acquisition department of the MPI that covers language acquisition corpora, longitudinal studies at different age groups and others also in many different languages.

Usually the included corpora are based on audio and video recordings, resources that take up much storage space so that the total archive already amounts to 11 TeraByte of data. Every year about 35 field trips are organized by MPI researchers and also the DOBES documentation teams are very active and are continuously integrating new data. In addition to this, from summer 2005 we will open the archive for

contributions from external projects that don't have these facilities themselves. We estimate that the amount of highly interesting language resources in the archive will increase by a factor 2 within 3 years. With these numbers in mind we started early to think of possibilities to automate procedures (Broeder 2004).

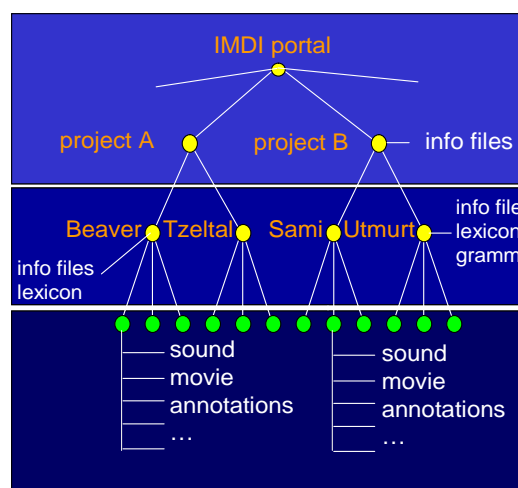


Figure 1 shows the IMDI archive/corpus organization principles. The yellow nodes are corpus nodes that can refer to information files of different sort. The green nodes represent detailed metadata descriptions of complex resources such as annotated media files or lexicons.

The whole archive is well organized according to the web-based IMDI catalogue principles (IMDI 2005). That is the archive structure is organized with the help of IMDI corpus nodes in sub-corpora and the resource bundles, lexicons or other relevant linguistic objects are described with the help of the IMDI metadata descriptions. Whether downloading the IMDI Browser, or using normal WWW browsers and selecting the MPI corpus portal (CORPORA 2005), one can browse and search in this MPI language archive domain without restriction since all metadata is open. The IMDI metadata descriptions include references to the resources (as is shown in figure 1) allowing suitable tools to be started from the browser. The structure of the corpora that is part of the archive is defined by the

projects or individual researchers, and of course they create the detailed metadata descriptions according to their insights.

Various checks that are carried out regularly guarantee a high degree of structural integrity of the archive. Where possible we transform different annotation and lexicon formats to one of a limited set of well-known formats such as EAF (ELAN 2005), CHAT (CHILDES 2005), Shoebox (SHOEBOX 2005) and LMF (LMF 2005). We can therefore achieve a high degree of format coherence in the archive. This all makes the MPI language resource archive an increasingly interesting source for linguistic analysis.

The IMDI metadata domain of described and registered language resources now covers contributions of more than 50 institutions, of which there are a number of well-known European language resource centers such as ELDA, BAS, Meertens, Lund University, Helsinki University, Florence University, ILC, ILSP, DFKI, various National Sign Language Communities, etc. These contributions were created within the ECHO and the INTERA projects. A few of these resource centers are participating in the DAM-LR project (Distributed Access Management for Language Resources). The tasks of this project have to not only create an integrated searchable and browsable metadata domain, but also to create an integrated storage and access management framework for different language resource archives. Also the possibility will be created for the different archives to refer to resources of each other through a generalized reference format independent of the actual physical location.

The goals of DAM-LR are in line with the requirements discussed by the international DELAMAN network that covers some of the major endangered languages and music archives which are collaborating (DELAMAN 2005). A necessary step to realize such a framework is the availability of a management and upload system for each individual archive.

In this paper we will describe the LAMUS system (Language Archive Management and Upload System) that is currently being developed at the MPI for Psycholinguistics. LAMUS can be seen as a content management system specialized for dealing with a large amount of (multimedia) language resources.

## 2. Archive vs. Digital Library

A very important question that has to be answered before creating a language resource archive concerns the nature of such an archive. Is it a traditional archive whose primary task is manage the long-term preservation of data. Alternatively, is it more of a digital library which has the primary task of giving access to the data and allowing its modification. Primary data such as recorded audio and video streams will not change. However, due to changing linguistic theories and insights, existing secondary data such as annotations and lexica will be subject to change. Further, it can be expected that new research groups will add additional secondary data such as new annotation tiers mostly in standoff format. Therefore the language archive, while it obviously should also deal with long-term preservation issues, is probably more related to the digital library. Language archives have to be concerned with both aspects.

Although multiple versions of resources must be supported, a language resource archive must not degrade to a researcher's or research group's workspace where all types of test and temporary versions are stored. Only resources that represent some final state of work and that are of high quality should be part of the archive. Since only formal criteria can be checked automatically a supervision mechanism has to be established to guarantee the quality of the archive content.

If new versions of an existing resource are uploaded, a versioning mechanism has to assure that old versions are not overwritten. Old versions need to be available a.o. to guarantee that existing references remain valid. Competing versions of resources, reflecting different scientific views on the data have to be available with equal ease so as not express a preference. It is clear however that the number of such "equal" versions must be limited, perhaps to a fixed number.

## 3. LAMUS

The considerations above are the motivation for the development of LAMUS, a web-based language archive management and upload system. In LAMUS a user can request one or several workspaces which are temporary areas where he can establish a new sub-corpus. Such a personal workspace has a limited time of existence and the amount of storage capacity requested is limited as well.

At the start a user specifies an existing part of a corpus in the archive which is then copied into the workspace. Then new data can be added or old data can be modified or deleted and if the user is satisfied the data in the workspace will be moved back into the archive.

More specifically the user has the following functions available in the workspace:

- corpus nodes can be created in a flexible way
- prefabricated session metadata descriptions can be inserted in those corpus nodes
- resources such as media files, annotations or lexica can be checked in
- unlinked resources can be easily linked with session descriptions
- final correctness and consistency checks can be carried out

Finally, when the checks have been carried out without errors, the user can upload the content of his workspace into the archive. This includes updating existing resources, if possible LAMUS will handle the versioning aspect transparently for the user.

Very important are the various checks carried out before moving the data into the archive, these checks should guard the consistency and coherence of the archive. LAMUS will allow an archive to define the accepted formats for the different linguistic data types. There are checks as to whether the uploaded resources have the correct formats, to assure that the expected archive coherence is guaranteed. Also checked is whether the metadata descriptions are well-formed according to the IMDI schema and whether the vocabularies used are coherent with IMDI. Finally, there is a check as to whether all new nodes and resources are linked correctly, i.e. unlinked and undescribed resources will not be accepted or integrated into the archive. The automatic checks cannot be extended to the linguistic encoding level for obvious reasons, although provisions are foreseen where resources are kept on hold in the workspace until some (human) validation agency has approved them.

After the user is satisfied with the situation in the workspace and the resources have been moved into the archive, LAMUS will ensure that the appropriate indexes are updated to

allow searching in metadata and annotation and lexicon content. LAMUS will also include enhanced features for archive management, i.e. the managers can execute the mentioned checks, run statistics on the archive and move data in a controlled way. Figure 2 shows a typical screenshot of the current version of LAMUS.

In summer 2005 a complete system will be delivered. LAMUS is based on Java Servlets and will be programmed such that it can be used and tuned by other institutions according to their wishes. It will also offer a number of APIs, for example, to be able to use the indexes for searching services. In doing so LAMUS can be seen as the basis for comprehensive archive exploitation. LAMUS will ensure that the language archive can be exploited and modified with new emerging web-based tools such as ANNEX and LEXUS (Berck 2005).

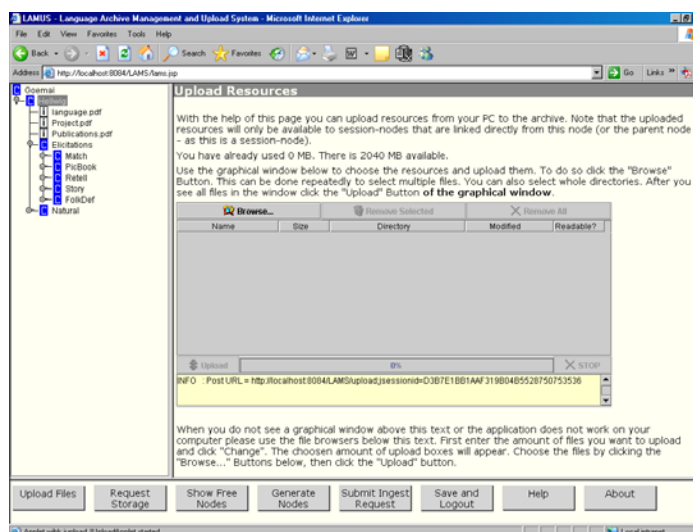


Figure 2 shows a typical screenshot of LAMUS.

#### 4. Access Management

The resources in a LAMUS private workspace are only accessible by the owner, i.e. all resources within the workspaces are closed per default. When they are moved from the workspace into the archive the resources will be equally closed for all users other than the donator. However, the responsible researcher can use the existing Access Management System (AMS) to define access policies (declarations to be signed) and set access rights for a group of users. Persons can be associated flexibly with groups.

AMS is built on top of the IMDI metadata infrastructure in so far as it allows the user to select a corpus node and define access policies for all resources that can be found under the chosen node with one command. It is also possible to distinguish between different resource types when specifying these rights. The commands will be entered in a relational database and then be expanded to htaccess files used by the Apache web server and Access Control Lists for accessing the archive via the file system. For media streaming, temporary URLs are created for the resource, since the streaming servers do not make use of the information in the htaccess files. AMS will be an integrated component of LAMUS.

#### 5. Future Developments

Within the DAM-LR project one of the goals is to create a unified access management infrastructure. The essential functions will be that (1) users will have one identity with which they can access the archives of the collaborating institutions and identification is done by specialized and shared services; (2) access rights will be associated with unique resource identifiers making it possible that all copies of a resource possibly stored by different archives share the same access policies; (3) authorization to access resources will be handled by specialized software; and (4) the collaborating institutions will set up mechanisms to copy each

other's resources on a large scale in order to increase the probability of long-term survival.

DAM-LR will be based on software that is currently being developed within the GRID community and related initiatives such as the Handle System (HANDLE 2005) to resolve unique identifiers, Shibboleth (SHIBBOLETH 2005) for access authorization, A-Select (A-SELECT 2005) for user identification and others more. These solutions will replace the current AMS solution in the long run and will also interact with LAMUS.

## References

- Broeder, D.; Brugman, H.; Oostdijk, N.H.J.; Wittenburg, P. (2004). "Towards dynamic corpora," In: N. Oostdijk, G. Kristofferssen and G. Sampson (eds.), Proceedings LREC 2004 Workshop on Compiling and Processing Spoken Language Corpora, 24 May 2004 Lisbon. 59-62..
- IMDI (2005). <http://www.mpi.nl/IMDI>
- CORPORA (2005). <http://corpus1.mpi.nl/BC/IMDI-corpora>
- ELAN (2005) <http://www.mpi.nl/tools/elan.html>
- CHILDES (2005) <http://childes.psy.cmu.edu>
- SHOEBOX (2005)  
<http://www.sil.org/computing/shoebox>
- LMF (2005) <http://www.tc37sc4.org/>
- DELAMAN (2005) <http://www.mpi.nl/delaman>
- Berck, P.; Russel, A.; Kemps-Snijders, M.; Wittenburg, P. (2005) ANNEX paper proposed to conference
- HANDLE (2005) <http://www.handle.net/>
- SHIBBOLETH (2005) <http://shibboleth.internet2.edu>
- A-SELECT (2005) <http://a-select.surfnet.nl/>