# The Language Archiving Technology domain

**Alexander Koenig, Jacquelijn Ringersma, Paul Trilsbeek**

Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands
{alexander.koenig,jacquelijn.ringersma,paul.trilsbeek}@mpi.nl

**Abstract**

The Max Planck Institute for Psycholinguistics (MPI) manages an archive of linguistic research data with a current size of almost 20 Terabytes. Apart from in-house researchers other projects also store their data in the archive, most notably the Documentation of Endangered Languages (DoBeS) projects. The archive is available online and can be accessed by anybody with Internet access. To be able to manage this large amount of data the MPI's technical group has developed a software suite called Language Archiving Technology (LAT) that on the one hand helps researchers and archive managers to manage the data and on the other hand helps users in enriching their primary data with additional layers. All the MPI software is Java-based and developed according to open source principles (GNU, 2007). All three major operating systems (Windows, Linux, MacOS) are supported and the software works similarly on all of them. As the archive is online, many of the tools, especially the ones for accessing the data, are browser based. Some of these browser-based tools make use of Adobe Flex to create nice-looking GUIs. The LAT suite is a complete set of management and enrichment tools, and given the interaction between the tools the result is a complete LAT software domain. Over the last 10 years, this domain has proven its functionality and use, and is being deployed to servers in other institutions. This deployment is an important step in getting the archived resources back to the members of the speech communities whose languages are documented. In the paper we give an overview of the tools of the LAT suite and we describe their functionality and role in the integrated process of archiving, management and enrichment of linguistic data.

**Keywords:** language archiving, linguistic annotation, lexicon tools

## 1. Introduction

Around 10 years ago the Max Planck Institute for Psycholinguistics set up a digital archive for linguistic resources (IMDI, 2009), an initiative which resulted in archiving the resources of the DoBeS, Documentation of Endangered Languages Project (Dobes, 2006). Currently the archive houses around 20 Terabytes of video, audio and textual resources in the language domain. These data do not just originate from DoBeS projects, but also from the MPI's researchers (MPI, 2009), who work in language acquisition, language comprehension and the relation between language, culture and cognition. Moreover, external institutions for instance, the Leiden University or the European Science Foundation (ESF) have stored data from their language projects in the MPI archive.

The archive serves a double purpose: on the one hand it has been set up as a means for the long term preservation of linguistic research data for future generations, and on the other hand it is already of use today for the exchange of research data between linguists or researchers from other scientific domains. Archiving is becoming more and more important, as many research organizations or funding agencies make it obligatory for researchers to store their data in a place where it is in principle accessible by the scientific community (Green et al., 2009). Moreover, the online availability of the archive also makes it possible for the members of the speech communities, whose languages are documented, to access the data. Whilst the archive is available online to meet the requirement of accessibility, in order to allow for legitimate interests of the researchers or the speech community and to comply with personality laws it is possible to restrict access to the resources.

With such an extensive archive and a large number of users accessing its data regularly, it is of major importance to ensure consistency and stability of the data there. Apart from a number of smaller tools to help the MPI archive managers controlling the archive and keeping possible problems in sight, there are three big cornerstones in archive management: IMDI, LAMUS and AMS. IMDI stands for ISLE metadata initiative, an initiative which developed a linguistic metadata standard that is now central to the MPI archive. The IMDI tools allow the researchers to easily create their metadata files describing the language resources (IMDI editor) and to find them back through the online IMDI browser and search engine. The Language Archive Management and Upload System (LAMUS) (Broeder et al., 2006) is an online tool which allows researchers to upload their data into the archive, and additionally performs a major role in the maintenance of the quality and consistency of the archive. Finally, the Access Management System (AMS) (Claus, 2004) permits researchers to set access restrictions of the resources in the archive, according to the specific needs of the researcher or the affected language communities.

Once archived, a set of archive exploitation tools offers access to the resources: ANNEX is the resource viewer, from which media resources (audio or video) can be viewed simultaneously with their transcriptions. TROVA is the content search engine, which allows users to perform complex searches in the archive's textual resources. Finally with IMEX, images and their specific metadata can be viewed in one thumbnail frame, from which specific images can be selected.

Besides the management tools and exploitation framework, the MPI has also developed a set of enrichment tools which can be of use for linguistic research or for making the resources more accessible to the speech communities. ELAN (Wittenburg et al., 2006) is by now a well established and well known annotation tool. LEXUS (Ringersma and Kemps-Snijders, 2007), is a lexicon tool, which can be used for the creation of online multi-media lexica, and with the closely related ontology tool ViCoS (Zinn, 2008) researchers and members of the speech communities may visualize culturally relevant conceptual spaces of related linguistic objects.

The integrated set of management and enrichment tools form the Language Archiving Technology (LAT) suite, which on the one hand helps researchers and archive managers to manage the data and on the other hand helps users to exploit their resources and enrich their primary data with additional layers. The LAT suite can also be of use for other scientific institutes, museums or cultural centers involved in documenting and archiving language. About two years ago the MPI started installing archives based on the LAT framework in various locations around the world (e.g. IIAP, Peru; CONICET, Argentina; IATSIS, Australia). The idea behind this initiative is to have regional archives in the proximity of the area where the linguistic resources are collected. This will facilitate access to the resources and create more local involvement and awareness towards the preservation of endangered languages and cultures (Trilsbeek et al., 2008). In the remaining sections of this paper we describe the tools of the LAT suite in more detail and we finish this paper by a section elaborating on the interaction between the tools.

## 2. Archive Management

### 2.1. IMDI

The MPI archive is hosting a large variety of data, both content and format wise and it is a requirement that finding the resources is easy. However, researchers are free to structure their corpora the way they want it (e.g. according to location of origin of the data or by thematic subject). Therefore, users who are unfamiliar with the structure of a specific corpus can experience difficulties in finding resources they are specifically looking for. For this purpose all data that is put into the archive has to have a metadata description file associated with it. This metadata description can then later be searched in a structured manner to comfortably find certain resources.

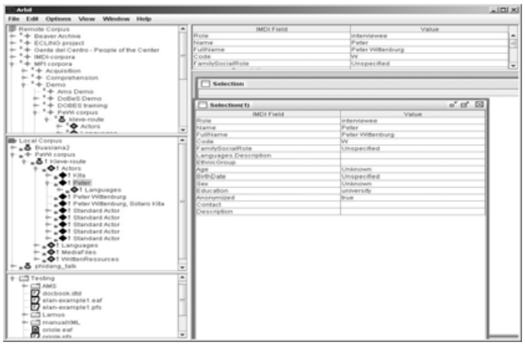
The metadata description of the resources is given in a so-called IMDI metadata file (Wittenburg, 2004), a file in an XML-based format, with a set of elements developed specifically for the description of linguistic resources. The set contains metadata elements on the context in which the data was collected (e.g. project, contact person, location) and elements which describe the content of the resources (e.g. language, genre, modality). Attached to the IMDI file are the linguistic resources pertaining to the same linguistic event: a set of media (audio, video or image) and written resources (transcription or analysis files).

IMDI is the metadata format by which the resources in the archive are described, and once archived the IMDI browser and a custom-made search tool can be used to find them. As the IMDI format is XML-based it is in principle possible to simply create and edit IMDI files by hand in any XML editor. But this is, of course, very arduous and error-prone, which is why the MPI has developed a graphical editor for working with IMDI files. The IMDI editor (IMDI, 2008) is built in such a way that it is not possible to create invalid IMDI files. It offers a structured way of entering values for the different metadata element and a controlled vocabulary for the fields which are most used. This behaviour helps in ensuring the integrity and consistency of the MPI archive at a very early stage.

The IMDI editor's counterpart on the consumer side is the IMDI browser. With this software it is possible to browse through all the metadata in the MPI archive, search for specific resources and – provided that the user has the necessary rights – view the resource itself. We do also supply other ways of accessing the data, for instance by facetted browsing or through portals designed for a specific linguistic community (Ringersma et al., forthcoming), but even these alternative ways of browsing the data in the archive use the IMDI metadata set or IMDI search as their backbone.

### 2.2. Arbil coming up

The IMDI editor is at this point in time almost ten years old and a lot has happened in software engineering. Therefore, instead of simply updating the IMDI editor to incorporate user suggestions and modern software architecture principles, developers at the MPI are working on an application that will replace the IMDI editor altogether in the near future. The new tool is called Archive Builder (Arbil) and it is already in its beta stage (see Figure 1).



| IMDI Field | Value |
|---|---|
| Role | interviewee |
| Name | Peter |
| FullName | Peter Wittenburg |
| Code | W |
| FamilySocialRole | Unspecified |

| IMDI Field | Value |
|---|---|
| Role | interviewee |
| Name | Peter |
| FullName | Peter Wittenburg |
| Code | W |
| FamilySocialRole | Unspecified |
| Languages Description | |
| EthnicGroup | |
| Age | Unknown |
| BirthDate | Unspecified |
| Sex | Unknown |
| Education | university |
| Anonymized | true |
| Contact | |
| Description | |

Figure 1: Arbil User Interface

The current version is downloadable from the LAT web site (Arbil, 2009), so it can be tested by interested readers. With Arbil, the IMDI concept and format did not change: the users get the same functionalities as the IMDI editor, but in a more user-friendly GUI, plus the new tool will also be closer integrated into the whole LAT software architecture, for example by providing a LAMUS interface to immediately upload newly created IMDI files to the archive. Like the IMDI editor, it is possible to use Arbil without any internet connection, since many of the MPI and DoBeS researchers are working in field conditions where they only have shaky internet access, if any access at all.

### 2.3. LAMUS

With LAMUS (Broeder et al., 2006) researchers can create the corpus structure that they find most suitable for their data. For this purpose LAMUS offers an easy user interface for adding and linking corpus nodes. Within LAMUS corpus nodes can be attributed a Name and a Description, two elements of the IMDI metadata set.

LAMUS is the only way for researchers to upload their IMDI metadata files and resources into the archive. Since the idea of the MPI archive is that data is stored persistent and consistent, it is important to control which type and format of data is ingested into the archive. LAMUS acts as the archive's gatekeeper. Because of its long term preservation goals the archive limits accepted resource file formats to what we call 'archivable formats' (Broeder et al., 2006). Those formats should be open and well-documented. Proprietary formats, like Microsoft Word, are explicitly excluded, one major reason being possible backward compatibility problems.

Moreover, while submitting the new data to the existing archive, LAMUS also generates persistent identifiers (URIDs) for the new files, thus creating persistency over time and independency of the resource location of each archived object. The MPI archive is using the Handle System to create, maintain and resolve URIDs (Handle, 2009).

Finally LAMUS sets some access rights on the new files. New files in the archive inherit access rights from the archive node to which they are attached; if no rights have been set on these higher nodes, the default setting is such that the newly uploaded resources are not accessible from the browser other than by the owner of the data (see also the next section on AMS).

### 2.4. AMS

There is a variety of reasons why a certain resource should not be accessible to everyone. Within the scope of this paper we do not wish to enter into the Intellectual Property Rights and access debate, but we do appreciate that, for instance privacy rights of the persons depicted in a video are of serious concern here. So without further taking a position in the current debate we do acknowledge the potential need for differentiated access. An elaborate Access Management System (Claus, 2004) that controls what users are able to see what data is available for the archived data.

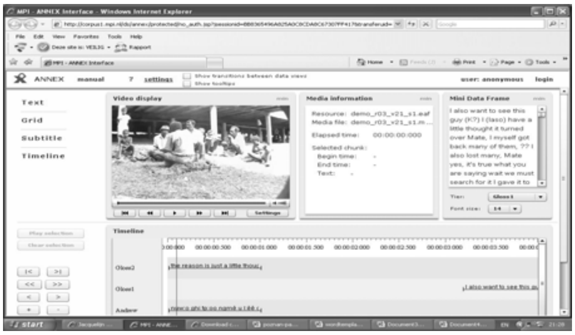
There are three levels of user access rights that can be assigned to a resource: the resource can be completely open to anybody, the resource can be open to all registered users who have signed a Code of Conduct or the resource can be open to a selected user or group of users only (Claus, 2004). Finally it is possible to close a resource completely. All these levels can be assigned to specific resources, to whole corpora or to sub-corpora. Different access rights can be set depending on whether a resource belonging to a certain corpus is a video, an image, an audio file or an annotation.

For the purpose of access rights, we differentiate between metadata, which in compliance with the Open Archives Initiative (OAI, 2009) are always open to everybody, and resources (e.g. audio and video files, pictures, annotations, etc.) for which access has to be set with AMS. By default a new resource is only readable by the person who put it in the archive. This person then has to set any further access rights explicitly. Additionally there is a priority-system in place that allows for complex rights scenarios, and the owner of a corpus can also assign certain roles to other users so that they can help them in managing access to their resources.

## 3. Archive Exploitation

### 3.1. ANNEX

ANNEX (Berck and Russel, 2006) is a Flex-based tool that views video resources in sync with their annotations. It uses the ELAN files to establish which media files are associated with a given set of annotations and how the annotation is synchronized with the media stream. In the ANNEX view users can select to view one or more of the annotation tiers and in different frame set-ups. Playing of selections and backwards playing are also within the ANNEX options. (see Figure 2). ANNEX can be triggered directly from the IMDI browser.



Resource: demo_r03_v21_s1.eaf
Media file: demo_r03_v21_s1.m...
Elapsed time: 00:00:00:000
Selected chunk:
Begin time: -
End time: -
Text: -

I also want to see this guy (K?) I (laso) have a little thought it turned over Mate, I myself got back many of them, ?? I also lost many, Mate yes, it's true what you are saying wait we must search for it I gave it to

Tier: Gloss1
Font size: 14

Figure 2: ANNEX

### 3.2. TROVA

TROVA is the archives annotation content search engine. TROVA allows the user to search over one or more annotation files. Users may perform simple keyword searches over all the content, or perform regular expression based searches over single or overlapping, related tiers, within annotations and over multiple annotations. Results can be viewed within the context of the searched item, and the size of the context can be specified between 1 to 6 annotations. The results can also be viewed in a frequency view, showing the frequency of the specified annotation. From the search result list, the annotation can be viewed in ANNEX.

## 4. Archive Enrichment

Primary resources collected in the field require analysis and elaboration in order to be able to study a documented language or a certain phenomenon in the language in more depth. Transcribing and translating the resource and adding morpho-syntactical annotations to it can be one of the first steps in this research process. For this we provide the researchers with ELAN (ELAN, 2009), a professional tool for the creation of complex annotations on video and audio resources. We further provide LEXUS (Ringersma and Kemps-Snijders, 2007), an online lexicon tool for the creation of multi-media dictionaries plus the related ViCoS (Zinn, 2008) tool, with which researchers, as well as members of the speech community, can create relations between items in the lexicon and thus create semantic or ontological networks which allow analyzing the culture of the documented language. In the following section we briefly describe these three enrichment tools of the LAT suite.

### 4.1. ELAN

ELAN (ELAN, 2009) is a tool for the manual creation of annotations to audio and/or video files. Annotations are stored in a Unicode-based XML format to guarantee wide cross-system compatibility and to ensure that the format is future-proof. Annotations in ELAN are basically text strings, either a sentence, word or gloss, a comment, translation or a description of any feature observed in the media (e.g. a description of certain gestures made). They can be created on multiple layers separately, which can be hierarchically interconnected. Usually annotations are directly time-aligned to the video or audio recording, but it is also possible to refer to other existing annotations instead. As some experimental set-ups require multiple different view-points, it is possible to link up to four different video files to the same annotation file. If a user is interested in the phonetic form of his recordings they can also link both a video and an audio file to the same ELAN file, so that the wave form can be seen at the same time as the actual video content. ELAN supports import to and export from a wide variety of other annotation formats (e.g. Shoebox/Toolbox or CHAT).

ELAN is configurable in a number of ways to enable the user to set it up in such a way that they feel comfortable with. Among others, videos can be displayed in the main window or in separate resizable windows, and annotations can be viewed in several different ways with each view being connected and synchronized to the media. Media playback is delegated to an existing media framework, like Windows Media Player, QuickTime or JMF (Java Media Framework). As a result a wide variety of audio and video formats is supported and high performance media playback can be achieved.

### 4.2. LEXUS

Creating a lexicon is especially interesting when documenting a less-known language. Since this is one of the major aims of the DoBeS projects, the MPI has developed a tool for the creation of online lexica, LEXUS (Ringersma and Kemps-Snijders, 2007). With LEXUS it is possible to create online dictionaries or thesauri and in addition it is also possible to attach multi-media fragments, like video, audio or images to the entries in the dictionary. The result is more than a simple lexicon, namely an attractive almost encyclopedic view on word lists collected by researchers and the speech community.

LEXUS is based on the Lexical Markup Framework (LMF) and therefore complies with the respective ISO standard for linguistic resources (LMF, 2006). Having said this, LEXUS also supports flexible lexicon structures and is compatible to the wide-spread Toolbox format.

Besides attaching multi-media fragments directly to the lexical entries, it is also possible to link from LEXUS to the resources in the archive. This makes LEXUS an attractive application to open up the resources in the archive in a different manner than through the IMDI browser, namely by the words in the word list. LEXUS allows web-based access and supports collaboration of several researchers or speech community members by implementing a workspace concept and a differentiated read and write rights system.

### 4.3. ViCoS

ViCoS (Zinn, 2008) is a tool designed for complementing lexical spaces (as created by LEXUS) with ontological spaces. With ViCoS users can create relations between the lexical entries in LEXUS, for which a set of 'universal' relation types (synonym, antonym, part-of relation, etc.) is provided. Lexicon specific relation types can be created in order to accommodate non-western based relations, which are of importance when documenting endangered languages and the cultures that come with these languages. Users then can define concepts they think are culturally relevant, and connect them to other concepts via a multitude of relation types. The result is a conceptual space of related items, which can be browsed from one concept to another (see Figure 3).

The tool is designed in close collaboration with the members of the speech community in the DoBeS project 'The Marquesan languages: 'Eo 'Enana and 'Eo 'Enata'. MPI developers try to develop it as user friendly as possible, the aim being to encourage members of the speech communities to use it to actively describe their own language and culture. This 'insider look' on a language can be preferable to the same work being done by linguists who can only view the language and especially the culture from the outside.

As ViCoS is based on LEXUS it is possible to link certain lexical items to media files in an archive so that users can get a better idea of how the word is used.

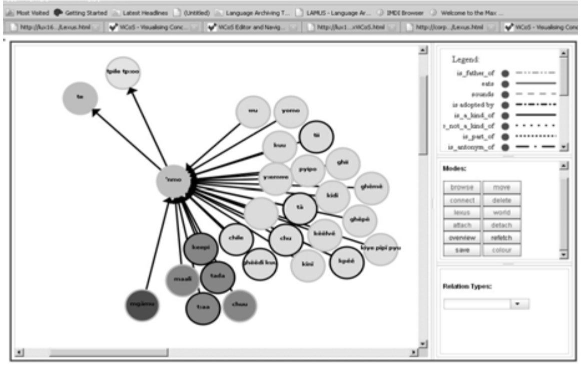


Figure 3: ViCoS browser

## 5. Integrating the LAT tools to one domain

We do make a point of all of the software being, at least on some level, interconnected with the archive, or with each other. Of course, the researchers' workflow starts by making recordings and digitizing them, but the way to the archive is describing the resources with IMDI metadata using the IMDI editor and uploading them through the gatekeeper LAMUS. The files are then moved into the archive while they are assigned a persistent and unique identifier. Next, the archive itself is the focal point of all connections. This is reflected in the main access software of the IMDI Browser. Apart from providing access to the archive by allowing browsing through it, a metadata search as well as a search inside of annotations can be started from the IMDI Browser. The former provides links to metadata files that can then be opened in the IMDI Browser again. Provided that the resources have been annotated with ELAN, the IMDI browser also opens TROVA, a powerful annotation search engine, and displays the results in ANNEX (Berck and Russel, 2006). The IMDI Browser has a further gateway function by providing easy access to the Access Management System (AMS) to set the rights for specific archive nodes, copora or sub-corpora. On the other side of the workflow AMS is automatically triggered by LAMUS to set certain default access rights for new additions to the archive.

The ontological tool ViCoS, and the lexicon tool LEXUS have been developed to be closely interconnected. The idea behind it being that the former depends on data stored in the latter to further visualize it. LEXUS is also able to interact with ANNEX for the visualization of annotated media files. In general, LEXUS (and ViCoS to a degree) can be seen as a new component in archiving, since it has possibilities of linking and interlinking linguistic as well as cultural concepts, thus creating a dense network of indigenous knowledge and concepts, which has not been achieved in conventional electronic language archiving so far.

## Acknowledgements

The authors of the paper are the archive managers in the MPI LAT team. We would like to acknowledge the tedious work of the developers in our team, who deserve all credit for the functioning and proper integration of the tools. In specific we would like to thank: Daan Broeder, Marc Kemps-Snijders, Han Sloetjes, Claus Zinn, Peter Withers, Mariano Gardellini and Thomas Koller.

## References

Arbil (2009) Arbil, Beta Version. Retrieved from: http://www.lat-mpi.eu/tools/arbil/ Access date: October 13, 2009.

Berck, P. and Russel, A. (2006) Annex - a web-based framework for exploiting annotated media resources. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC, 2006)*

Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P. and Wittenburg, P. (2006) LAMUS: The Language Archive Management and Upload System. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

Claus, A. (2004). Access management system. Language Archive Newsletter, 1(2), 5

DoBeS (2006) Retrieved from: http://www.mpi.nl/dobes/ Access date: October 13, 2009.

ELAN (2008) Retrieved from: http://www.lat-mpi.eu/tools/elan/ Access date: October 13, 2009.

GNU (2007) General Public License - Retrieved from: http://www.gnu.org/copyleft/gpl.html. Access date: October 13, 2009.

Green, A., Macdonald, S. and Rice, R. (2009) Policy-making for Research Data in Repositories: A Guide. Data Share Project. Retrieved from: http://www.disc-uk.org/docs/guide.pdf Access date: October 13, 2009.

Handle (2009) Corporation for National Research Initiatives- Retrieved from: http://www.handle.net/ Access date: October 13, 2009.

IMDI (2009) Domain of IMDI Described Corpora - Retrieved from: 1839/00-0000-0000-0000-0000-4 Access date: October 13, 2009.

IMDI Editor (2008) Retrieved from: http://www.lat-mpi.eu/tools/imdi/editor/ Access date: October 13, 2009.

LMF (2006) The Lexical Markup Framework. Retrieved from http://www.lexicalmarkupframework.org/ Access date: October 13, 2009.

MPI (2009) Retrieved from: http://www.mpi.nl/ Access date: October 13, 2009.

OAI Open Archives Initiative (2009) Retrived from: http://www.openarchives.org Access date: October 13, 2009.

Ringersma, J. and Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme & R. van Son (Eds.), Proceedings of Interspeech 2007 (pp. 1529-1532). Adelaide: Causal Productions.

Ringersma, J., Zinn C. and Koenig A. (accepted). Eureka! User friendly access to the MPI linguistic data archive. Workshop "Usability Aspects of Hypermedia Systems" October 1st, 2009, University of Potsdam/Germany

Trilsbeek, P., Broeder, D., Van Valkenhoef, T., & Wittenburg, P. (2008). A grid of regional language archives. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Wittenburg, P. (2004). The IMDI metadata concept. In S. F. Ferreira (Ed.), Workingmaterial on Building the LR&E Roadmap: Joint COCOSDA and ICCWLRE Meeting, (LREC2004). Paris: ELRA - European Language Resources Association.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

Zinn, C. (2008). Conceptual spaces in ViCoS. In S. Bechhofer et al. (Eds.), The semantic web: Research and applications (pp. 890-894). Berlin: Springer.