LR Archive at the MPI

Report
for the
Archive Advisory Board

13.11.2005

Peter Wittenburg
Paul Trilsbeek

Content

1. Introduction

In this report we will only present the essentials that have to do with the stability and accessibility of the archive. This may be sufficient to raise questions of all sorts that can be raised during the coming meeting of the AAB.

The document describes the essentials guiding the construction and maintenance of the language resource archive housed at the MPI for Psycholinguistics. The main archive contains the following sub-archives

- the MPI archive
- the DOBES archive
- the ESF corpus
- the CGN corpus
- others (various other contributions)

The sub-archives are physically part of the same archive, but logically and with respect to access policies they are different. The notion of sub-archives is recursive, i.e. the contributions from different DOBES teams are also separated in the same way.

Further, the MPI has decided to open its archive for contributions from other projects or even individuals. This has to do with the fact that according to D. Schüller about 80% of our recordings about cultures and languages are endangered. So, we have to help digitizing important material and integrating it into our repository. However, sub-archives bound to certain agreements and quality standards can only be extended by involving the corresponding boards. For other contributions separate areas will be established.

Further, we house a portal for metadata that is based on IMDI. Any site that wants to participate will be registered. Currently the IMDI portal harvests metadata from several institutions in Europe.

This report will discuss the following aspects:

- archiving concept
- archiving principles
- state of the archive
- long-term preservation issues
- access management
- access tools

2. Archiving

2.1 Concept

There is an ongoing debate about the tasks of modern digital archives. On the one hand traditional archives were focusing on long-term preservation of physical objects and therefore had to restrict the access. On the other hand modern digital archives know that in general copying can be done without loss of information[1] which means that almost unlimited access can be given. In addition, web-based technologies allow users to not only access material, but also to enrich the content (commentary, corrections, extensions, relations, etc). To not affect the stored data, however, the enrichments have to be stored independently of the original data and new versions may not overwrite the older versions.

[1] There may be loss of information when format conversion is being carried out, in particular for compressed formats.

Due to these fundamental changes, the MPI archivists see it as their major tasks to take care of all three aspects

- taking care of long-term preservation
- giving access to the content to authorized persons via a set of methods
- allow authorized persons to update and enrich the content

The digital archive at the MPI is seen as a living repository the content of which is dynamically changing without that the original data is being affected.

2.2 Principles

The archive should follow a number of principles (current practice is indicated by cursive comments):

1. The physical storage layer (disks, servers, directory paths) and the logical layer (metadata) should be separated, since the physical layer will be changed frequently due to migration and copying activities. It is the logical metadata layer that is defined by the depositors, makes use of the domain terminology and will remain stable. This means that for any object in the archive a metadata description has to be created and that the links have to be maintained. The metadata domain is open for everyone.
   The MPI is making use of the IMDI set and this requirement is fulfilled. The IMDI set is currently part of the ISO standardization process, i.e. it will be maintained for some years.
2. The archival objects get a Unique Resource ID (URID[2]) which is shared by all its instances (due to copying activities). Access rights are associated with the URIDs to assure that they are the same for all instances.
   The MPI has developed a concept based on the Handle System within the DAM-LR project and will move towards URIDs in the beginning of 2006.
3. The content should be in open, well-documented and widely used formats. For textual data open schemas should describe the structure of the documents. Where possible generic schemas such as LMF (see below) should be the basis.
   The MPI has different strategies based on the agreements:
   - *For the DOBES archive strong agreements hold, i.e. only the following encoding/structuring standards are accepted for the archive objects: JPEG, TIFF, PNG, MPEG2, Linear PCM (>=48 kHz, >=16bit), UNICODE, XML, Plain Text, HTML (restricted), PDF[3]. For presentation purposes HTML, MPEG1/4 and MP3 are accepted.*
   - *The MPI archive was set up with less restrictions and it is already now the case that there are formats that have to be converted to make the content accessible, i.e. conversion effort is done to convert MediaTagger, Transcriber, the "old" ESF format, CHAT, SHOEBOX, EXCEL and WORD files. The archive also includes various plain text files created with unknown tools the structure of which has not been checked.*
   - *The archive will convert CHAT and SHOEBOX files to XML/UNICODE, but also store the original files.*
   - *EAF can be seen as a fairly generic schema for complex annotations. Yet there is no standard being worked out that satisfies all requirements. LMF is a proposal for a generic lexicon schema being worked out currently within ISO. Both are used as key formats to represent annotations and lexica within the MPI archive.*

[2] A URID can be compared with an ISBN number in the domain of publications.

[3] The mass of textual resources in the DOBES programme from the first 15 teams will come during the coming months. Much conversion work is expected, since the teams use various tools such as WORD for their work. Evangelization efforts to rely on tools operating with schema-based XML formats were not successful. Efficiency and convenience have highest priority.

4. Long-term preservation is focusing on bit-stream preservation and is done by migration to new storage and server technology at regular time intervals and by extensive copying to other trustful institutions.
   The MPI currently generates 7 copies of all its archival data at different sites, each of them having a clear migration policy (see below).
5. Access Management has to be guided by policies (code of conduct, copyright statements, usage declarations) that have to be accepted by the persons involved. It has to be efficient and password protection is seen as being sufficient.
   The MPI has set up an efficient web-based access management system. It is based on the metadata infrastructure and includes a delegation mechanism. It also asks the user to accept the Code of Conduct or comparable declarations. The usage request and declaration steps agreed upon in the DOBES project have to be integrated.
6. Authentication and Authorization systems have to be separate.
   The MPI currently uses a special database for authentication and HT-Access for authorization. For in-house users to the MPI sub-archive currently an access via the file system is possible. This feature, however, will have to be taken out of operation soon. In the DAM-LR project it will step over to LDAP for authentication and continue to use HT-Access for authorization. Shibboleth will probably be introduced for distributed authorization.
7. The archive has to support the bundling of resources according to their relations at different layers. At the language layer it must be possible to integrate resources that describe the language as a whole (such as a lexicon) and at the resource layer it must be possible to bundle for example video or audio recordings with their annotations. There are other aspects of bundling in between.
   The MPI allows the user to draw all these object relations with the help of the IMDI infrastructure.
8. The archive objects have to be stored in a neutral way and they may not be encapsulated, i.e. once access rights are given, the user must be able to access any individual object in its open format.
   The MPI follows this principle. Encapsulation (into database systems or HTML for example) is only done for internal optimization purposes (search engines) and presentation purposes. The original resources are not changed.
9. Enrichments may not affect the existing archival content.
   The MPI has developed the web-based LAMUS system so that authorized users can extend or upgrade archival content. Currently, a suitable versioning mechanism is being developed. Also web-based archive content enrichment frameworks are being developed. They will always upgrade archival content by making use of LAMUS as archive gatekeeper. Commentary and arbitrary content relations will have to be stored separately.

3. State of the Archive

3.1 Storage and Access System

The archive storage and access infrastructure was completely renewed in 2004/2005. It is now based on the following components:

- SAM-FS as hierarchical storage management system and core of the system serving for generating local copies also
- 2 redundant SUN servers (state-of-the-art 4 processor servers) with sufficient memory capacity (16 GB)
- a SCSI-based RAID system of 4 TB as fast cache
- a S-ASA-based RAID system of 16 TB as slow cache for media
- a Tape Library of 55 TB
- a Fibre Channel interconnect
- several LINUX servers to support all types of access, to run the applications and to store indexes

The SAM-FS software is very reliable and generates immediately two copies of each new resource integrated to the archive. It operates on the physical layer and is under the control of the system managers.

The following major software components are used:

- Solaris
- Linux Suse
- Apache
- Tomcat
- Postgres
- Java for all major developments
- Perl for conversion and management programs
- AFS Client

3.2 Content Overview

The MPI archive has three major divisions that have to do with the work process:

- The **Media Archive** contains all metadata described objects. It currently has a size of 4.4 TeraByte. The DOBES archive is part of this archive and contains 333 GigaByte. This is that part of the archive that is searchable and accessible by authorized users.
- The **digitization team** maintains for the DOBES a repository that contains 4.9 TeraByte and for all other contributions a repository with currently 6 TeraByte. These contributions are not yet metadata described and not yet organized according to the agreed principles.

Further, we can give the following overview about sessions and type of resources in the accessible **media archive**:

	MD sessions	video files	audio files	photo	other media	textual files	sub-types
MPI	18524	14085	5131	7774	1315	13979	365 EAF, 2377 CHAT, 5580 MediaTagger, 3568 PlainText/Shoebox, 1589 others
DOBES	1396	1043	1250	63	20	205	46 EAF, 85 Shoebox, 72 others
Dutch Spoken Corpus	12767		12767			41832	to be converted to EAF
Dutch Bilingual Database	874		191			714	CHAT, EAF
ECHO Sign Language	168	296				181	in EAF
ESF corpus	994	546				1775	in CHAT
Total	34723	15970	19339	7837		58686	136555 objects

Comments: EAF is schema-based XML; Photos in HTML files are not counted; some files claim to be SHOEBOX and CHAT files, however, they are not correct; since there are no explicit schemas there is no chance of verification

Much conversion has to be carried out in particular in the MPI part of the archive to come closer to a coherent representation based on acceptable formats. The Shoebox files in the DOBES archive will be converted when the final versions will be deposited.

In total the archive housed at the MPI requires a capacity of 15.3 TeraBytes which is an increase of about 4 TeraByte during the last year.

It is obvious that this state is not satisfying. It indicates that the researchers are failing to describe their data in time with the risk that after some period no one knows what the data is about etc. The MPI recently decided that minimal metadata has to be generated by every researcher and that resources have to be integrated into the Media Archive. We are waiting on measures by the directorate to address these issues.
For the DOBES part we assume that the teams that will finish soon (15) will completely organize their material before the end of the funding period.

3.3 Long-Term Preservation

The MPI for Psycholinguistics will upgrade its technology at regular intervals (4 to 10 years dependent on the component) and it stores two copies in its building at two different locations.

In addition the following steps have been taken:

- A complete copy is maintained at the Computer Centre of the Max-Planck-Society in Garching (RZG). The RZG has an exchange program with the Leibniz computer center at the university of Munich so that actually two copies are maintained. The copy is upgraded dynamically by using the Andrew File System software.
- A complete copy is maintained at the Computer Centre of the Max-Planck-Society in Göttingen (GWDG). The GWDG has an exchange program with the computer center at the university hospital in Göttingen so that actually two copies are maintained. The copy is upgraded dynamically by using the rsync program.
- A complete copy of the DOBES archive is maintained at the MPI for evolutionary Anthropology in Leipzig. The copy is upgraded dynamically by using the Rsync program.

With all three sites written documents have been exchanged that specify the rules of behavior.

In addition, the following measures have been taken:

- The Max-Planck-Society takes over an institutional guarantee for 50 years that the data archived at its computer centers will be taken care of.
- The MPI is busy to create a DELAMAN[4] domain in which major archives agree on documents that will allow us to distribute data all over the world while retaining the right to define access options (procedures, policies, rights) by the original site.
- The MPI currently is testing out the necessary kind of distributed scenario in the DAM-LR[5] project together with SOAS (U London), U Lund, Institute for Dutch Lexicology Leiden.

As already indicated we see the necessity to unify the number of formats in the archive, i.e. much conversion has to be done to improve the coherence.

4. Archive Upload and Access

4.1 Access Management

Two persons at the MPI are responsible for managing the archive:

- Paul Trilsbeek
- Roman Skiba

[4] Digital Endangered Languages and Music Archive Network
[5] Distributed Access Management for Language Resources

These archive managers interact with the system managers of the MPI who define and control the specifications at the physical layer and the tool developers. No operation may be carried out on the archive without permission of the archive managers.

It should be mentioned that the archive claims the right to archive the deposited data and that it claims copyright on behalf of the creators. In special agreements with every depositor it is specified what the policies will be with respect to granting access to data that is not open; it could be the case that the depositor specifies that the archive managers have the right to give access to their data[6].

An **Access Management System** was developed that has the following features:

- The two archive managers at the MPI have full access right and the right to define access rights. They are bound to adhering to the laws and the Code of Conduct.
- Only the archive managers have the right to delete files (only to be used in very special cases).
- Depositors have the right to access the deposited data.
- They can delegate all rights (except deletion) to other persons in particular to the depositors (in the DOBES case the responsible researcher per team).
- These persons can further delegate the rights in their sub-archive to trusted persons.
- Per sub-archive policies can be defined. In the case of DOBES every user has to accept the Code of Conduct[7].
- The authorized person can select an archive node (IMDI node) in his/her sub-archive and specify access rights for all resources of a specific type (video, audio, texts) with one command.
- The rights are stored in a database and extended to records per resource[8] in the HT-Access file.
- The HT-Access file is scanned by the Apache web-server when a person tries to access files.
- For the streaming server that give access to media data a temporary link file is created that is used by the streaming application.
- Also for content searching (currently in the test phase) access rights will be checked.

This AMS will be replaced partly by a new system that is prepared to operate in a distributed scenario as well. Here LDAP, Shibboleth, Handle System and additional components will play a role. The MPI will inform the AAB (and others) about steps to be carried out to get advice.

Currently, most of the resources in the archive are protected, i.e. not accessible. This is true for all DOBES data.

4.2 Archive Management

Many programs and scripts have been developed to check the state of the archive:

- check of dead links and link stability
- check relation between physical and logical structure
- check of metadata correctness including controlled vocabularies and congruence with file extensions
- check of format coherence including parsers for CHAT and SHOEBOX files
- create various statistics
- create a metadata crawler that can be extended with various modules

[6] This happened already several times with requests from journalists.
[7] This will be extended to other declarations to be accepted such as about the usage of a resource.
[8] As was indicated, metadata is open per definition.

The functionality of these scripts and programs is being integrated to LAMUS (see below) stepwise to make LAMUS a comprehensive management tool.

4.3 Metadata Access

The IMDI "standard" stabilized and is widely accepted as one of the two relevant metadata standards in the area of language resources. It was worked out based on suggestions from field linguists and language engineers within the European ISLE project and work on it was continued in the INTERA, ECHO and LIRICS projects. The most recent IMDI version 3.0.3 is a product of broad discussions based on the experience with earlier versions. The whole IMDI domain is based on open distributed XML files, i.e. the IMDI files are in an archivable format and everyone can build his own services and does not have to rely on MPI tools.

- The **IMDI Editor** is a very professional and user friendly tool with several options to increase the efficiency. The editor supports constraints and controlled vocabularies; therefore it is the only tool that guarantees that proper IMDI files are generated. The basic development can be seen as finished. Currently, further debugging is carried out and occasionally functionality is added.
- The **IMDI browser** is a valuable tool since standard web-browsers still do not support schema-based XML domains such as IMDI. The browser is ready and has many options such as simple and complex attribute oriented search. It is currently the best tool for browsing and searching in IMDI-based metadata domains. It will always have some advanced functionality which other options cannot offer.
- The IMDI domain is browsable via **HTTP web-browsers** since IMDI files are on the fly being translated to HTML pages. This transformation is ready and operational.
- Structured and unstructured **search options** are available both within the IMDI browser and within the HTML environment. Structured search options offer the element set and the known vocabulary elements. Unstructured search is Google-like and also scans prose text fields in the metadata descriptions.
- The IMDI files are also searchable via **Google**.[9]
- Recently, the MPI started using **Google Earth** to allow geographic browsing. This is in development, but will be released soon.
- The **DC/OLAC bridge** allows OLAC (DC) service providers to harvest IMDI files and include them into their searchable domain. The new bridge was realized as a product of a close interaction between E-Meld and MPI specialists. For metadata harvesting the OAI PMH is offered, but the open XML files can be accessed and harvested directly, of course.
- The **TreeCopier** allows to easily download complete sub-archives including the resources to create fully functional new archive instances.
- The **BatchModifier** function allows the user to carry out changes in the IMDI files that are below the selected node.
- **Manuals** and **training course materials** are available for all major tools.

4.4 Resource Upload

Resource upload (integration) into the archive can be done via two paths:

- manually by the archive managers
- by using the LAMUS gatekeeper software

[9] A new applet with optimized user navigation has been created. Yet it is not possible to go to the appropriate archive node which was supported in the old version.

Manual integration requires much attention from the archive managers in particular to keep the links stable and to take care of format consistence (as much as possible/necessary). We have recognized that this is becoming increasingly impossible given the amount of resources to be integrated.

Therefore, the LAMUS software was developed which is now in an extensive test phase for several months already including users from the MPI. It has the following features for authorized users:

- web-based operation
- request of a work space that is normally given for a period of time
- specification of an accepted upload node (archive anchor)
- extend and manipulate the corpus structure
- upload metadata descriptions
- upload any type of resources
- create a linked sub-archive in the workspace and integrate this into the archive
- before integration a number of checks are carried out as indicated in 5.2
- LAMUS can handle a (hierarchical) list of accepted file types and parsers can be associated with types, i.e. not accepted files are rejected
- per default the integrated resources are not accessible except for the depositor
- it generates content indexes when new resources are integrated to enable fast searching

It is intended to have a configurable list of accepted file types so that every archive can define its rules for content and keep it flexible. Hierarchies are important, since complex file types such as SHOEBOX files or complex lexica are in general made of a bunch of closely related files.

Yet, LAMUS does not have a suitable versioning mechanism. We are busy to develop such an extension, so that updating resources can be handled without problems. LAMUS has to be extended to offer APIs so that tools supporting complex resources and web-based annotation options such as ANNEX and LEXUS (see below) can directly be used to upgrade archival content without bypassing the gatekeeper functionality.

In future the LAMUS program has to be extended to support URIDs and the DAM-LR scenario (as well as all the other tools).

4.5 Simple Resource Access

For the following access discussions we assume that access rights have been given as described above.

As indicated all resources are available directly in the specified formats. Users can

- download them and use local tools such as ELAN
- directly visualizing/playing them via HTTP access and browser plug-ins

These basic methods are not very helpful and can only be recommended for quick inspection. More important is the possibility to download complete sub-archives, i.e. selecting an archive node and downloading all resources below that node including the metadata descriptions and nodes. In doing so a new sub-archive can be created at another site that is immediately be operational. The TreeCopier allows the authorized user to do so.

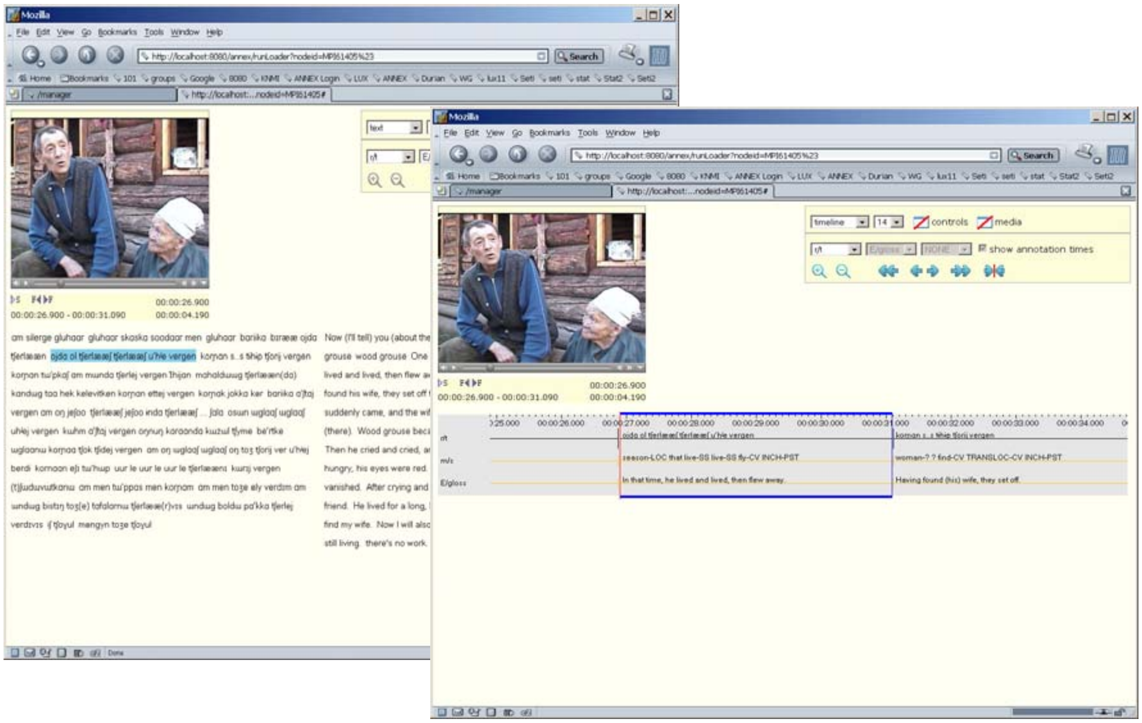
4.6 Complex Resource Access

It should be stressed that the principle of neutral direct access to the resources may not be broken, i.e. the shells that will be discussed in the following have to be seen as optional.

Everyone should be in the state to develop his/her own access tools. Nevertheless, we have to make access attractive and simple.

Web-browsers only allow to select and visualize/play a single resource. However, most of the relevant resources in language archives are of a complex type: (1) Annotated media files can consist of several video and audio recordings and multiple files with annotations. (2) Multimedia lexica can also contain of lexical entries that have photos and multimedia clips incorporated. Therefore, web-based frameworks are needed to access such complex types and analyze and compare the contents.

ANNEX (Annotation Exploitation Framework)



00:00:26.900 - 00:00:31.090 00:00:26.900 00:00:04.190

am silerge gluhaar gluhaar skaska soodaar men gluhaar banika bıraææ ojda tʃerlææn ojda ol tʃerlææʃ tʃerlææʃ u'hie vergen korŋan s...s ŧhip tʃorj vergen korŋan tu'pkaʃ am munda tʃerlej vergen Ihijan mahalduwg tʃerlææn(da) kandug taa hek kelevitken korŋan ettej vergen korŋak jokka ker banika a'ʃtaj vergen am oŋ jeʃoo tʃerlææʃ jeʃoo inda tʃerlææʃ ... ʃala osun wglaaʃ wglaaʃ uhiej vergen kuhm a'ʃtaj vergen arynuŋ karaanda kuzul tʃyme be'rtke wglaanw korŋaa tʃok tʃidej vergen am oŋ wglaaʃ wglaaʃ oŋ toʒ tʃorj ver u'hiej berdi kornaan eʃı tu'hup uur le uur le uur le tʃerlææns kurŋ vergen (t)ʃuduvutkanw am men tu'ppas men korŋam am men toʒe ely verdım am undug bistıŋ toʒ(e) tafalamw tʃerlææ(r)vıs undug boldu pa'kka tʃerlej verdıvıs iʃ tʃoyul mengyn toʒe tʃoyul

Now (I'll tell) you (about the ... grouse wood grouse One ... lived and lived, then flew a... found his wife, they set off ... suddenly came, and the wi... (there). Wood grouse bec... Then he cried and cried, a... hungry, his eyes were red. ... vanished. After crying and ... friend. He lived for a long, ... find my wife. Now I will als... still living. there's no work.

00:00:26.900 - 00:00:31.090 00:00:26.900 00:00:04.190

	00:00:27.000 – 00:00:31.000	00:00:31.000 –
rt	ojda ol tʃerlææʃ tʃerlææʃ u'hie vergen	korŋan s...s ŧhip tʃorj vergen
m/s	season-LOC that live-SS live-SS fly-CV INCH-PST	woman-? ? find-CV TRANSLOC-CV INCH-PST
Elgloss	In that time, he lived and lived, then flew away.	Having found (his) wife, they set off.

Mozilla
File Edit View Go Bookmarks Tools Window Help
http://localhost:8080/annex/runLoader?nodeid=MPI61405%23
Search
Home Bookmarks 101 groups Google 8080 KNMI ANNEX Login LUX ANNEX Durian WG lux11 Seti seti stat Stat2 Set2
/manager
http://localhost:...nodeid=MPI61405#
text
timeline
14
controls
media
rt
Elgloss
NONE
show annotation times
00:00:25.000
00:00:26.000
00:00:27.000
00:00:28.000
00:00:29.000
00:00:30.000
00:00:31.000
00:00:32.000
00:00:33.000
00:00:34.000
Done

These screenshots show views that are implemented with ANNEX. It gives a number of controls and selection possibilities and shows the annotations in different views such as the time-line view.

ANNEX is one of the components of a web-based exploitation framework being developed. It is intended to give flexible access to annotated media files via the web. Ideally one would expect to find ELAN[10]-like functionality. The first version of ANNEX will only be an exploitation framework, i.e. it is not meant to add or modify annotations. Exact and smooth media handling via the web is still a difficult problem.

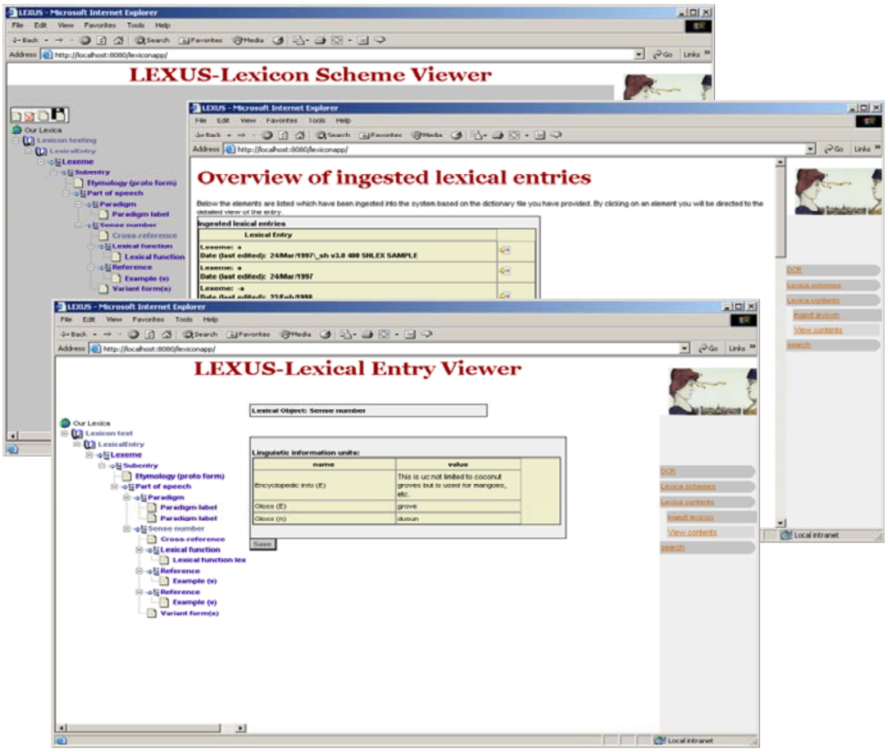
ANNEX works together with a smart search engine that allows the user to look for certain patterns of texts in certain annotation tiers and combination of those. It also has an unstructured search option. It allows to search across several resources that may come from different depositors, terminological differences are handled at this moment via simple search lists[11].

[10] ELAN is a local tool to annotate and analyze annotated media files.

[11] An extension to using ontological knowledge is in development.

ANNEX makes use of media streaming facilities which are based on MPEG4. It will officially be launched in January 2006. Currently, ANNEX does not have features to allow the manipulation and extension of the existing annotations. This will be added in 2006.

LEXUS (Flexible Lexicon Tool)
LEXUS is meant to be a full-fledged lexicon creation, manipulation and analysis tool. It is based on the new and emerging LMF (Lexicon Markup Framework) standard being developed within ISO TC37/SC4 and on the excellent ideas of Shoebox. The idea behind LMF is that the lexicon model should be as flexible as possible to represent all different possible lexicon structures. In particular in the area of endangered languages analyses by Wittenburg, Peters and Drude, 2002, see appendix A) showed that all lexicons have different structures dependent on the languages and intentions of the researcher. LMF can be used together with a data category registry as being developed also within the ISO framework to achieve semantic interoperability. LEXUS supports currently two registries: the ISO and the Shoebox MDF categories[12].


LEXUS-Lexicon Scheme Viewer

Overview of ingested lexical entries

Below the elements are listed which have been ingested into the system based on the dictionary file you have provided. By clicking on an element you will be directed to the detailed view of the entry.

Ingested lexical entries

Lexical Entry
Lexeme: a Date (last edited): 24/Mar/1997_sh v3.0 400 SHLEX SAMPLE
Lexeme: a Date (last edited): 24/Mar/1997
Lexeme: -a Date (last edited): 23/Feb/1998

LEXUS-Lexical Entry Viewer

Lexical Object: Sense number

Linguistic information units:

name	value
Encyclopedic info (E)	This is uc:not limited to coconut groves but is used for mangoes, etc.
Gloss (E)	grove
Gloss (n)	dusun

The screenshots are taken from an earlier version of the LEXUS tool. It allows to create and modify lexicon structures and analyze, create and modify lexical entries.

LEXUS has at this moment the following features:

- specification of an arbitrary lexicon structure that is suitable to the needs;
- manipulation of a given structure;
- import of structures from Shoebox and CHAT files or extract structures from given XML files;
- export to Shoebox and CHAT files;
- selection of lexical attributes from data category registries that have a format compliant to the ISO DCR or from the Shoebox MDF categories;
- visualization of lexicon structures in several ways;
- visualization of lexical entries in several ways;
- add, delete or manipulate lexical entries

[12]

- simple search across several lexica;
- export of a lexicon and lexicon structure to an LMF compliant XML file;
- inclusion of text, video, audio or image files;
- operation as a web-based tool, but also as a local tool on PC or MAC;
- first simple ways to merge two lexica;
- etc

LEXUS has and will have a number of features that go beyond the possibilities of Shoebox such as media support, interaction with multi-media frameworks, web-based operation, synchronization of lexicon versions, searching across several lexica etc. However, it will also have a few missing functions such as for example parsing capabilities. Therefore, a smooth import/export relation with Shoebox is important. The user interface will be modified stepwise to find a compromise between simplicity and functionality.

Lexus was officially launched in 2005 and is available for testing and operation. An interested user can receive request a workspace. As indicated, LEXUS has already features to allow users to manipulate structure and content via the web. Yet, it does not interact directly with LAMUS. This will be added in 2006.

LANA (Language Archive Access)
LANA is a web-based interface to integrate all web-based components created at the MPI such as ANNEX, LEXUS and LAMUS. The functionality that will be realized stepwise in 2006 and 2007 can be summarized as follows:

- manipulation of annotation contents via ANNEX
- ontology support for ANNEX and LEXUS to facilitate cross-corpus work
- combination of metadata and content searching
- integrated operation between LEXUS and ANNEX
- integrated operation between LAMUS as gatekeeper and LEXUS and ANNEX as content manipulation frameworks
- web-based commentary and tagging (which is a kind of annotation)
- web-based relation drawing to allow users to create type relations on content level

5. Final Statement

The responsible persons for the MPI archive will inform the Archive Advisory Board about changes with respect to the major principles, the basic hardware and software components, the access management principles and the access tools. The responsible persons are:

• Peter Wittenburg	Head Technical Department
• Paul Trilsbeek, Roman Skiba	Archive Managers

Any disagreement from AAB members with respect to proposals will be taken seriously. If no comments will be made from AAB members within a limited period of time[13] agreement is assumed.

[13] In general four weeks is seen as being sufficient.