LAMUS – the Language Archive Management and Upload System

Daan Broeder, Andreas Claus, Freddy Offenga, Romuald Skiba, Paul Trilsbeek, Peter Wittenburg

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
{daan.broeder,freddy.offenga,romuald.skiba,paul.trilsbeek,peter.wittenburg}@mpi.nl

Abstract

LAMUS is a web-based service that allows researchers to deposit their language resources into a language resources archive. It was developed at the MPI for Psycholinguistics for stricter control of the archive coherence and consistency and allowing wider use of the archiving facilities without increasing the workload for archive and corpus managers. LAMUS is based on the use of IMDI metadata standard for language resources and offers metadata search and browsing over the archive.

1. Introduction

The language resource archive at the MPI for Psycholinguistics stores digital language resources from the institute's groups for acquisition, gesture and cognition studies and also houses the corpora of related projects such as DOBES [1] and DBD [2]. Due to newer and increasingly cheaper technologies for recording, digitization and storage the archive has now reached a staggering, at least for the domain of language studies, total of 15 TB comprised out of 150000 individual objects. This amount is ever increasing due to the 60 expeditions per year from MPI and DOBES teams that bring back an average of 30 tapes.

The archive contains a large variety of different linguistic data types, i.e., (annotated) media recordings or text sequences, lexica, series of photos, field notes, sketch grammars, ethnological notes etc. Most of the archive is comprised by digitized recordings: both audio and video and the files containing the transcriptions and analysis. Next to these, there is IMDI metadata describing the individual resources as also their mutual relationships and dependencies. The relationships between resources are embodied by embedded links in the metadata [3,4].

The institute used to be able to manage the archive with a sizeable group of corpus managers that took care of the whole process of archiving from digitizing the media tapes, moving the files into the archive in suitable linguistic determined groupings and adding the metadata (provided by the depositors). Also the corpus managers were responsible for updating existing content and maintaining specified access policies. In fact they were and partly still are the only interface between the researcher/depositor and the archive.

2. Changing the Data Ingestion Workflow

Some time ago we deliberated the possibility of a different workflow for ingesting resources into the archive, one that relies on more involvement of the depositor, using modern web-based services integrated closely with existing archive access services and procedures. There are several arguments for changing to such a system, that we call LAMUS (Language Archive Management and Upload System.

2.1. Increasing costs.

The enhanced possibilities for recording, digitization and storage also increase the workload for corpus managers. There is no balancing force against the creation of raw unanalyzed material that is stored in the archive for possible future processing and analysis. This can be worthwhile data nevertheless but some minimal description and analysis of this data should be available before accepting it into the archive.

2.2. Using Depositor Knowledge

The depositor is the best qualified person to determine the way his resources should be integrated into the archive. However he may be not the best qualified person to deal with the physical realities of the archive like file systems and setting access permission. Therefore corpus managers performed this task, but needed much interaction making it questionable if it really saved the depositor that much time.

2.3. Remote Archiving Service

In the age of the internet and web based services we see a huge potential for offering remote archiving services. Many projects are already distributed i.e. have researchers with affiliations of different universities and institutes. Using a remote archiving service they will be able to ingest their data in a central archive profiting from essential services as guaranteed backup and access.

2.4. Stricter Checks

In the old system much depends on the knowledge of the corpus managers concerning archive policies such as what are the policies like resource naming, acceptable formats etc. At the MPI there is a big reliance on student-assistant work for corpus manager tasks, who tend to have very short-term contracts and often makes for less than perfect knowledge transfer. An automatic system that monitors the type, formats and interrelations of the ingested data can be a better gatekeeper and guarantee the archives coherence and consistency. This enormous change is indicated in figure 1.

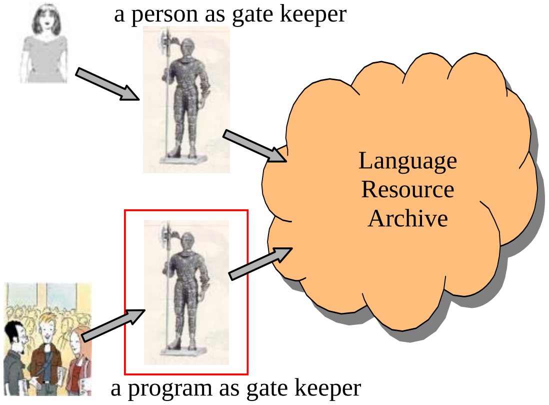


Figure 1 indicates the basic problem each language resource archive is confronted with. While until now individuals approached the archive manager as gate keeper, to take care of integrating objects into the archive, now we are confronted with a much larger group of depositors and much more data. This requires software that takes over the role of the archive manager as gatekeeper.

2.5. Maximizing Deposition

According to an overview made on request by the UNESCO [5] a large amount (80%) of very important data about cultures and languages are in danger to be lost forever, if they will not be handed over to powerful enough digital archives. Of course, the MPI and the DOBES archives feel the necessity to open their gate for contributions from third persons. However, this can only be managed when the load on the archive managers will not increase, i.e., a software controlled upload option is a prerequisite to solve the huge problem of loosing data.

It may be clear that the above arguments are related only to the archive data ingestion process, they are independent from those for advocating web-based services for access and utilization of the archived data. We think the case for the last has already been proven and we won't go into that here except to describe these services as a complement to the ingestion system where needed.

3. Depositor Guided Data Ingestion

As functional requirements for LAMUS we considered the existing archive workflow for data ingestion and listed the actions that presumably can be managed by the depositor such as:

- Uploading and naming individual resources (media, annotations, information files)
- Specifying the metadata and mutual relations for and between resources .e.g. IMDI resource bundles.
- Creating relevant linguistic groupings for the data, naming and arranging the material in sub-corpora.
- Specifying the access rights and policies for the deposited material. Required functionality is the possibility to specify access for specific known groups and users as also specifying requirements for users to first sign a code of conduct before they can access the material.
- Downloading individual resources or whole sub-corpora for the purpose of updating or local analysis and uploading it to its original location in the archive.

The system would then augment the depositor actions by:

- Carrying out many checks to guarantee consistency and coherence with the archiving rules (accepted formats etc) when uploading resources.
- Carrying out typical management operations such as updating databases and indexes and creating statistics.

4. Infrastructure requirements for LAMUS

Since these upload and management services are a part of the total archive infrastructure they also have to implement a number of requirements related to infrastructure:

4.1. Universal Resource Identifiers (URIDs)

The MPI's archive has decided to introduce stable identifiers for its resources. The problems pertaining to the use of URLs are well known [6], therefore a decision was made to use the Handle System (HS) of the CNRI [7] to provide a highly available service for resolving URIDs to actual URLs. The HS is well known in the library community. Adopting it will guarantee stable references from non-local resources (stand-off annotations) and publications.

4.2. Versioning.

The "stable identifier" issue from the previous point makes no sense if the resource itself is modified. Therefore, the original resource should never be deleted and always be accessible (although it need not be immediately). Also when we have a reference to a resource, we would like to be able to have access to older and newer versions if they exist. So when new resources are uploaded and the depositor specifies they are to replace existing ones, LAMUS needs to first move the old resources to the archive's "attic". Discussions on the visibility in views on the archive of the old versions are complicated, but for the moment we have decided on allowing only access to older versions on the basis of a direct reference to it or via a reference to another version of it. This divides the "viewable" archive in two dimensions: (1) the set of all latest versions of all objects in the archive and (2) on the basis of a selected archive objects we have access to its older versions.

4.3. Distributed Authentication

Although the MPI archive aims at self sufficiency, we are part of different projects and organizations such as DELAMAN [8] and DAM-LR [9] that aim at cooperation at different levels. Firstly, the cooperating archives share a group of users that would like to access resources housed at different places without maintaining different user accounts. Therefore the archives should accept each others authentication of users. An accepted solution for this is the Shibboleth system [10] that will be used within DAM-LR. Secondly, the cooperating archives can host copies of each others data for safety, preservation and availability reasons.

4.4. Modularity

The MPI has offered LAMUS to be installed at other interested archive organizations. Since the needs and available resources vary considerably amongst archives, for instance not every archive is prepared to maintain a URID infra-structure, LAMUS is set up in such a fashion that such functionality is an optional addition.

5. LAMUS Functions and Workflow

LAMUS is a completely web-based service that can be used by all main-stream web-browsers. Its main functions and usual steps in the workflow are:

1) Allow a user to apply for an account (if none has been issued yet) by specifying his identity, affiliation, what kind of data is going to be uploaded and where the data should be linked to in the logical organization scheme. This request has to be approved by a corpus-manager, and in some cases it may be necessary to ask the advice or permission of boards.

2) Once this request has been accepted the user is able to create one or more workspaces where the researcher can upload resources and metadata descriptions and do all sorts of manipulations as long as the maximum allowed storage capacity is not overwritten. The user can specify relations between all uploaded components in the workspace to create a proper corpus. At any step the user can check the state of his work.

3) When finished for the day, the user can suspend the workspace and reconnect to it another time and continue working.

4) Once the user has finished all uploading and manipulations, he can submit a request to move the data into the archive and at that moment further checks will be carried out to guarantee the compliance with archive standards and rules.

5) When data is moved into the archive, it will also move into the domain of URID addressable objects and therefore all embedded URLs need to be replaced by URIDs. LAMUS will also take care of necessary versioning operations.

6) Relevant databases will be immediately updated so that all ingested resources are visible for everybody via the metadata browsing and search infrastructure. In our archive metadata is open, however, access to resources themselves is barred by default unless the user has specified otherwise by setting special rules for this corpus.

7) Changing the default access permissions can be done by using efficient means, i.e., the user can choose the top node of a sub-corpus and specify in one single operation that all annotations thereof should be open to the world. An access management system component is part of LAMUS and its functionality has already been described elsewhere [11].

8) LAMUS will also automatically update index files that support fast metadata and content search, although the latter is restricted to text formats for which suitable parsers are available. Content search on annotations is supported by ANNEX [12], a web-tool developed at the MPI for viewing annotation files. The upload of resources will also trigger the update of a large index that will speed up content searching.

Once the resources and metadata have been ingested in the archive they can be downloaded either individually or as a “local” corpus by special tools. The resources and metadata in the downloaded corpus keep all their interrelations by adapting all embedded links to the new situation. LAMUS allows for such “local” corpora to be uploaded again into the archive and recognizes the existing embedded links, this minimizes the construction phase in the workspace. The workflow is shown in Figure2. LAMUS is shown as a shell around the archive allowing users to create workspaces initialized with existing data from the archive (1), uploading new data into the workspace (2) and finally copying the data from the workspace into the archive (3). Figures 3 and 4 show (part of) the LAMUS user-interface.

6. Conclusions

A core LAMUS system has been operational with increasing functionality since 2005 [13]. The experiences of the users, both from the MPI and external users have been guiding the further development. Currently we are implementing the URID and versioning additions which we plan to finish this year.

The use of LAMUS is thought also to be able to increase awareness at the depositors side about the resources to be deposited. Think for instance of a Shoebox [14] lexicon that comes along with a structure file and even language files, without it the data is not complete. However the researchers is not always aware of this and in our archive we found very little shoebox files accompanied by such structure files. If possible, we see possibilities for LAMUS to guide the depositor here and explicitly demand if such structure files are also available if he uploads a shoebox file.

There are many more of these cases where LAMUS should be aware of the possible or even required existence of auxiliary files.

A necessary extension of LAMUS not described in this paper, is to make programming APIs available that allow advanced tools to directly interact with the archive without going through the phases of creating workspaces and explicitly uploading resources. For instance the lexicon tool LEXUS [15] that has its own workspace and guidance mechanism for resource creation. It needs to use LAMUS functionality directly to ingest the lexica in the archive.

To test the portability of LAMUS we recently installed it at Lund University. This was an excellent exercise to see that within half a day the complete infrastructure including some corpora from Lund University was up and running [16]. The corpus can be viewed via Internet and the researchers at Lund University can upload new resources. A training course was held to show users and archive managers how to work with LAMUS.

7. References

[1] http://www.mpi.nl/DOBES
[2] http://www.ru.nl/dbd/start.html
[3] http://www.mpi.nl/IMDI
[4] Wittenburg, P., Peters, W., Broeder, D. (2002). *Metadata Proposals for Corpora and Lexica.* In M. Roriguez Ganzalez & C. Paz Suarez Araujo (eds.), Proceedings of the 3rd International Conference on Language Resources and Evaluation. Paris: European Language Resource Association. pp 1321-1326
[5] D.Schüller: http://www.mpi.nl/LAN/vol_01/lan_v01_n03.pdf
[6] Erickson, John. "Digital Object Identifier", In McGraw-Hill Yearbook of Science & Technology 2003.
[7] http://www.handle.net/
[8] http://www.delaman.org/
[9] http://www.mpi.nl/dam-lr
[10] http://shibboleth.internet2.edu/
[11] A. Claus, Access Management System. Language Archive Newsletter, 1(1), 5
[12] http://www.mpi.nl/annex/
[13] Claus, A ,Wittenburg, P ,Broeder. D. (2005) *Language Management and Upload System.* 2nd Language Technology Conference L&T 2005, Posen.
[14] http://www.sil.org/computing/shoebox/
[15] http://www.mpi.nl/lexus
[16] http://dam-lr.sol.lu.se/ds/imdi_browser/

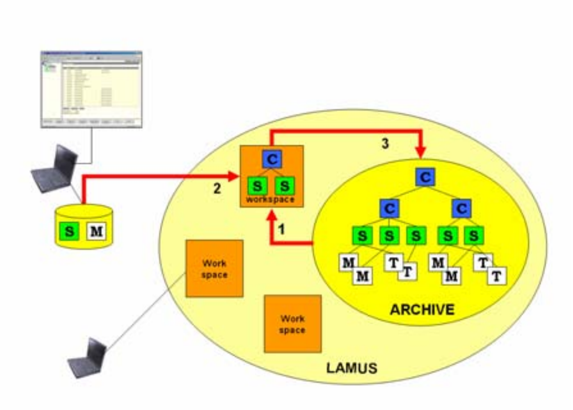
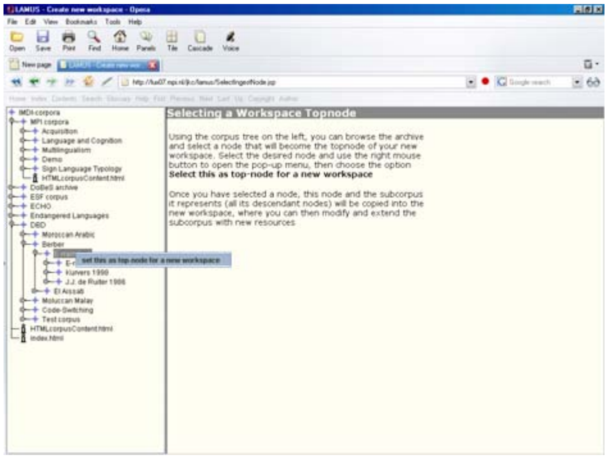


Figure 2 indicates LAMUS the basic workflow. New resources can be uploaded from a notebook or another archive into the workspace and from there into the archive. A user can also copy archive resources into the workspace for further processing and then upload them again as new versions. The icons stand for [M] media and [T] textual resources, corpus metadata [C] describes the linguistic groupings and resource metadata [S] describes resources and their interrelations.

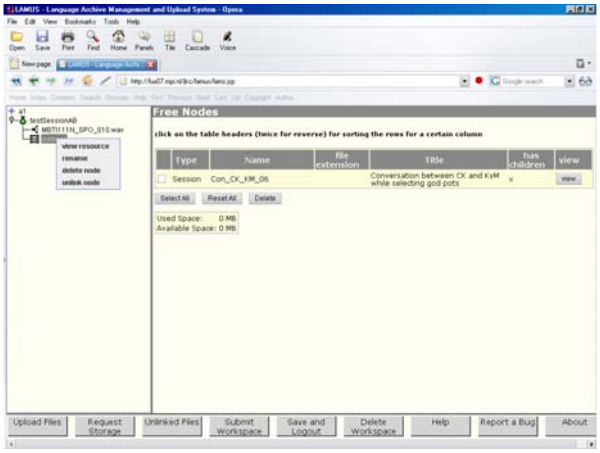


Selecting a Workspace Topnode

Using the corpus tree on the left, you can browse the archive and select a node that will become the topnode of your new workspace. Select the desired node and use the right mouse button to open the pop-up menu, then choose the option **Select this as top-node for a new workspace**

Once you have selected a node, this node and the subcorpus it represents (all its descendant nodes) will be copied into the new workspace, where you can then modify and extend the subcorpus with new resources

Figure 3 how part of the archive is selected to initialize a new workspace.



Free Nodes

click on the table headers (twice for reverse) for sorting the rows for a certain column

	Type	Name	file extension	Title	has children	view
□	Session	Con_CK_KM_06		Conversation between CK and KyM while selecting god pots	x	view

Used Space: 0 MB
Available Space: 0 MB

Figure 4 shows a view at a LAMUS workspace (left) and at the list of uploaded files (right).