

February 2009

Relevance

CLARIN community

for all communities

# Long-term Preservation and Access

## What is it?

Taking care of the long-term preservation of the covered resources is one of the most important tasks of a research infrastructure. While tools have a comparatively short life-span, and standards that guide integration and interoperability need to be supported for a number of decades, the data resources of the domain, such as recordings, annotations, lexica etc. need to be preserved for hundreds of years, if not for eternity. Linguistic resources document the state of cultures and languages in a time of extreme changes. CLARIN wants to address these needs by pointing to those centres that are able to give specific long-term preservation and access services.

## What is it for?

We can assume that

- researchers want to analyze, for example, the social climate change by studying language usage over time;
- politicians want to derive decisions from the sociological knowledge embedded in language resources;
- future generations want to understand their identity and may want to go back to original, less mixed forms of languages

Various initiatives, such as the ESFRI Preservation Task Force and the Alliance for Permanent Access (APA), have pointed out that the preservation of the record

of research is currently one of the major points of concern. This is, in particular, true of the linguistic domain, knowing that from the currently spoken 6500 languages, one is dying almost every week, and that even the main languages are subject to continuous changes due to globalization. However, the individual linguist, and even most of the research institutes in general, need to be freed from the burden of taking care of permanent accessibility with all its aspects. As has been indicated, an economy of scale factor can reduce the enormous costs.

## Who can use it?

- Preserving data resources is associated with costs for maintaining the bit streams, and especially for taking care of the data curation (format and encoding migration, consistency etc.) and yet there are no general policies. Therefore each of the CLARIN centres that are offering preservation services at this moment has its individual restrictions, deposit rules and archiving principles.
- This implies that all researchers and institutes that would like to deposit resources need to get in contact with those that are offering such a service. In principle the services are open for all interested researchers who have language resources.

A more systematic approach to achieve a better coverage and enabling automatic copying strategies will be necessary in future. Initiatives such as APA need to define policies that will work across disciplines.



## When can it be used?

CLARIN has already identified a few institutions that are capable of offering preservation and curation services. Depending on the material, these services can already now be used, i.e. these institutions can be contacted. To achieve more general services, solutions at political level need to be worked out.

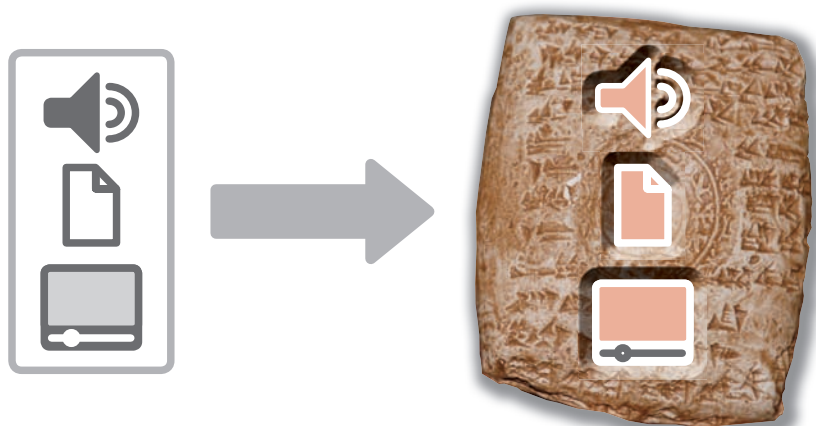
## How does it work?

CLARIN will mention on its web-sites the services that the different institutes are offering and refer to the corresponding information pages. From each institution offering preservation services, CLARIN will expect clear statements about the details and the services being offered, in particular the type of guarantees that are given and the type of deposit agreements. CLARIN

will also motivate these institutions to participate in a self-assessment process about their services, as it is offered, for example, by DANS with their “**Data Seal of Approval**” principles.

The interested user can get information about the deposit conditions at the corresponding web-sites, and dependent on the result, they can contact an expert from the selected institution. If an agreement can be achieved, users can deposit their resources according to the required procedures. One of the criteria for deposits in CLARIN certified institutions is the provision of proper metadata by the depositor, which guarantees that the resources are visible to the community at large.

CLARIN is currently looking for a more systematic approach to long term preservation together with other initiatives.



## Who is responsible?

CLARIN can only act as a mediator, for example, taking care that the details of the services being offered are made as explicit as possible and that self-assessments are carried out. Each institution needs to maintain a website providing information about the service, and suggest a contact for interested researchers. Centre's offering long-term preservation and access services are a mixture of Type A and B centres.

## Whom to contact?

For the CLARIN research initiative work package 2 is taking responsibility to maintain a web site with service providers (<http://www.clarin.eu/??>). In case of questions contact:  
Dieter van Uytvanck (MPI): [dieter.vanuytvanck@mpi.nl](mailto:dieter.vanuytvanck@mpi.nl)

## Where to find more information?

The official CLARIN web-site is the source of all information:

CLARIN: <http://www.clarin.eu>

CLARIN Centres: <http://www.clarin.eu/specification-documents>

CLARIN Centre types: <http://www.clarin.eu/specification-documents>

ESFRI Document: [ftp://ftp.cordis.europa.eu/pub/esfri/docs/digital\\_repositories\\_working\\_group.pdf](ftp://ftp.cordis.europa.eu/pub/esfri/docs/digital_repositories_working_group.pdf)

APA Document: [http://www.alliancepermanentaccess.eu/documenten%5CAlliance2008conference\\_report.pdf](http://www.alliancepermanentaccess.eu/documenten%5CAlliance2008conference_report.pdf)

e-IRG Document: [http://www.e-irg.eu/images/stories/reports/2006-report\\_e-irg\\_tf-sei.pdf](http://www.e-irg.eu/images/stories/reports/2006-report_e-irg_tf-sei.pdf)

Responsible for the content:

**Dieter van Uytvanck**  
**MPI for Psycholinguistics**  
**Wundtlaan 2, 6525 XD Nijmegen, NL**  
**Website: [www.clarin.eu](http://www.clarin.eu)**  
**Email: [dieter.vanuytvanck@mpi.nl](mailto:dieter.vanuytvanck@mpi.nl)**