Nomination form
International Memory of the World Register

Selected data collections of the world's language diversity at The Language Archive

ID Code [2014-84]

1.0 Summary (max 200 words)

The nominated heritage offers and preserves a unique sample of the world's linguistic and cultural diversity. It represents a landmark for language documentation in terms of quality of content and archiving infrastructure.

The nominated holding at the Language Archive (TLA) consists of 64 digital collections with audio-visual and textual resources documenting 102 languages and cultures around the world, recorded and annotated for providing high-quality representative samples of the respective languages in their natural context, digitally prepared and archived between 2000 and 2014. They depict little known and understudied communities and will be the irreplaceable major or only future record of the respective languages / cultures, created by world leading specialists with active involvement by these speech communities.

Care has been taken that the collections are and remain available online, by creating technology that ensures their digital long-term preservation conforming and contributing to emerging international standards for digital archiving. Ethical requirements such as individual's privacy and protection of cultural sensitive materials have been respected.

The inclusion in the UNESCO Memory of the World will enhance TLA's capacity to raise funding for further enhancing the holding's usability for research, for exposition materials, and in language maintenance activities.

2.1 Name of nominator (person or organization)

The Language Archive, Max Planck Institute for Psycholinguistics, Max-Planck-Society (TLA / MPI-PL / MPG, Nijmegen, Netherlands),
with:

De Koninklijke Nederlandse Akademie van Wetenschappen (KNAW, Netherlands)

Die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW, Germany)

Die VolkswagenStiftung / Volkswagen Foundation (VWS, Germany)

2.2 Relationship to the nominated documentary heritage

TLA/MPI-PL/MPG: Responsible institutions for hosting and preserving the heritage

Main funding partners of TLA (KNAW, BBAW) and main funder of the development of the infrastructure and of projects that contributed to the heritage holding (VWS)

2.3 Contact person(s) *(to provide information on nomination)*

Sebastian Drude

2.4 Contact details

Name	Address	
Sebastian Drude Paul Trilsbeek	Wundtlaan 1, NL–6525 XD Nijmegen, Netherlands	
Telephone	Facsimile	Email
+31 24 3521 470	+31 24 3521 213	Sebastian.Drude@mpi.nl
+31 24 3521 203		Paul.Trilsbeek@mpi.nl

3.0 Identity and description of the documentary heritage

3.1 Name and identification details of the items being nominated

If inscribed, the exact title and institution(s) to appear on the certificate should be given

Selected data collections of the world's language diversity at The Language Archive

*The nominated heritage is a holding, consisting of 64 individual digital collections with audio-visual and textual resources at **The Language Archive of the Max-Planck-Institute for Psycholinguistics, Nijmegen, Netherlands, belonging to the Max-Planck-Society**. The Language Archive is supported by the Max-Planck-Society, by the **Berlin-Brandenburg Academy of Sciences** and by the **Royal Dutch Academy of Sciences**, and has received substantial funding by the **Volkswagen Foundation** for the work on setting up the nominated holding and its infrastructure. Therefore, these institutions participate in the nomination.*

The holding comprises all collections as present in March 2014 in this archive for which the right holders have agreed so far for the collection to be included in the UNESCO Memory of the World programme.

The holding concerns digital resources from 102 languages and cultures around the world, recorded and processed so as to provide high-quality samples of the respective languages being used in their natural cultural context. The collections have been curated, digitally prepared, compiled and archived at The Language Archive between 2000 and 2014 (see attached list for details).

The collections contain natural speech – monologues such as myths and other narratives or oral traditions, oral history, explanations, personal stories, etc., but also dialogues (conversations etc.), or elicitation/interviews on language structure or cultural context.

The holding and its collections were carefully created and compiled according to the maxims of "language documentation", i.e. to cover not only certain isolated aspects of the language and culture, but to provide a sample of a broad spectrum of genres and language usage situations. In particular, the major part of the collections has been created in the context of the international research programme "Documentation of Endangered Languages" (DoBeS). Many of the collections also document important cultural events and everyday culture.

3.4 History/provenance

The collections in the holding here nominated were collected under the lead of a scientist or a team of scientists, and with active involvement of the members of the respective speech communities. Worldwide leading scientists from dozens of advanced universities and academic institutions, in particular the Max-Planck-Institute for Psycholinguistics, were involved in the creation of these materials. The active involvement by the speech communities also in design and compilation of the collections is particularly strong in the case of the more recent collections.

Many collections here included have been digitally compiled by researchers and local co-workers around the world in the years 2000–2014, in the context of the DoBeS programme (from German DOkumentation BEdrohter Sprachen, with funding by the Volkswagen Foundation, which is not a subsidiary of the respective company, cf. http://dobes.mpi.nl). The DoBeS programme had been envisaged since the early 1990ies, and the Universal

Declaration of Linguistic Rights ('Barcelona Declaration') and the publication of the first edition of UNESCO's Atlas of the World's Languages in Danger, both in 1996, have had significant influence on the establishment of the programme in 2000.

Other collections in this nomination have been created by researchers of the Max-Planck-Institute for Psycholinguistics in the context of research projects on the relation between language, culture, and cognition, since the 1990ies. Finally, some relevant collections have been collected in the context of other unrelated research or cultural heritage projects.

All collections of this holding have been carefully curated and archived at The Language Archive (TLA) of the Max-Planck-Institute for Psycholinguistics over the last 15 years by specialized TLA staff in cooperation with the researchers and project members. Several collections also contain materials which had been recorded before this period.

4.0 Legal information

4.1 Owner of the documentary heritage (name and contact details)

This varies for the individual collections: Please see separate list in Appendix A.

Name	Address	
Telephone	**Facsimile**	**Email**

4.2 Custodian of the documentary heritage (name and contact details if different from the owner)

Name	Address	
Sebastian Drude	*MPI-PL, Wundtlaan 1* *NL – 6525 XD Nijmegen*	
Telephone	**Facsimile**	**Email**
+31 24 3521 470	*+31 24 3521 213*	*Sebastian.Drude@mpi.nl, TLA@mpi.nl*

4.3 Legal status

[For this kind of material, it is hard to differentiate between ownership / legal status and copyright. We treat these issues together here, and refer in section 4.5 to this section.]

Ownership and copyright are both shared between the research teams and the speech communities. The database right on the holding as a whole lies with The Language Archive at the Max-Planck-Institute for Psycholinguistics which belongs to the Max-Planck-Society (MPG), established together with the Dutch Royal Academy of Sciences (KNAW) and the Berlin-Brandenburg Academy of Sciences (BBAW).

Care has been taken to obtain informed consent from the speakers and speech communities. Still, due to the often very unequal degree of literacy and familiarity with digital and network technology between the research teams and speech communities, there are often no formal written agreements (but, for instance, the explanations and negotiations themselves have been recorded). In the collections in the DoBeS framework, the funding was granted under the condition that as much as possible of the resources would be made easily available for further research wherever ethical concerns and privacy of speakers allow. Respect for the speech community, their culture and their concerns have been of paramount importance in the DoBeS framework and at the MPI for Psycholinguistics. How these concerns have been attended varied according to the settings and conditions of the individual projects in the context of which

the individual collections of the holding have been created.

We understand that all participants in the creation, compilation and archiving of each of the collections here nominated participate in ownership and copyright, which includes the respective speech community (collectively, where appropriate, and the individual members that participated) as well as the researcher / research team involved in the recording and compilation of the materials, and the archiving institution.

The precise formal legal situation of the collections may vary, also due to the international setting – speakers, researchers, funders, the archive, and users may all belong to different countries and institutions, each with their own laws and rules.

4.4 Accessibility

All materials are on-line and reachable via The Language Archive's web-interface via http://tla.mpi.nl. Each individual file (e.g., an audio or video recording, or a text file with annotation) and each "session" (thematic bundle of files) have persistent identifiers that ensure that the materials remain accessible even if the physical or logical location or the names of the material change. The Language Archive has set standards of reference in terms of building a sound infrastructure for maintaining the accessibility for the future. Bit-stream preservation has been guaranteed for 50 years by the Max-Planck-Society.

The resources have primarily been created and archived for scientific use, besides use by the speaker communities and their descendants.

All the metadata-descriptions are always open and freely available on-line for anyone.

Each of the (audio-visual and textual) resources themselves has currently one of four possible access levels:

1) *open: can be freely accessed without logging in*
2) *open when logged in: registering and agreeing to a code of conduct (see below) is required (this is the default)*
3) *open for certain users (restricted): access needs to be requested*
4) *open only for owners (closed): access cannot be requested (exceptional, total <50)*

The distribution of the resources over these access levels is as indicated in the attached list (Appendix A).

Access levels (3) and (4) are necessary for some resources in order to be able to allow for an embargo period where researchers have the opportunity to publish results before opening the data, or in order to be able to respect the privacy of speakers and for protecting culturally sensitive materials. Importantly, the (members of the) speech communities themselves have been consulted when determining the access rights; they have to agree in order to open any resources.

Generally, the policy of The Language Archive is to encourage the owners to do whatever they can to make as much as possible of the material as freely accessible as possible, and overall we see an increase in freely accessible resources (levels 1 and 2).

TLA is going to introduce soon this year an additional user level (2a) – "open for academic users", where TLA checks the academic affiliation, and academic users will only have to agree to a usage agreement and the code of conduct (for DoBeS materials, http://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf) in order to access the material. Then most of the material that is now in access level (3) is to be re-assigned to the new access level unless specific pertinent reasons prevent this for individual resources. This is expected to further enhance the accessibility of a large part of the collections for academic usage.

4.5 Copyright status

As said above under "legal status", the copyright is shared between members of the speech communities or usually the speech communities collectively, and the researchers.

The Language Archive (TLA) and the Max-Planck-Institute for Psycholinguistics do not claim any copyright to the content, but TLA takes responsibility for the database as a whole.

5.0 Assessment against the selection criteria

5.1 Authenticity.

Authenticity of a digital holding applies to two levels – that of the content and that of the digital medium. The digital medium is particularly intricate, as in principle any access to any digital resource creates a new copy of the 'stream' of bits and bytes (sequence of zeros and ones).

Technically, the collections consist of many digital (computer) files, i.e. individual (pieces of) audio and video recordings or individual pictures, but also textual files (annotation, metadata, etc.). The multimedia-files are high fidelity copies of the original recordings. Older recordings have been transferred into digital formats using professional equipment and software. This has generally been done at The Language Archive itself, rarely at the researcher's institutions and only exceptionally at other places (as in the case of very old media carriers such as steel wire). Newer recordings have been made directly in digital formats ('borne digital') and were handed over to TLA on digital carriers. The original data carriers, both analogue and digital, are always returned to their owners; TLA is not a physical but exclusively a digital archive.

If necessary, the digital files have then been converted into current standard formats appropriate for archiving and dissemination, being as faithful as possible to the original recordings, again, using professional equipment and software. If in the future file format conversions are necessary because archival formats are getting outdated and superseded, TLA will preserve the original archived files together with their new, derived versions.

There are several layers of digital infrastructure, of software and hardware, between the visualization of the materials on a screen and the physical carriers where the bits and bytes are stored. At the MPI in Nijmegen, TLA uses a modern hierarchical storage system where materials that have not been accessed in a while are available on high-performance hard-disks while often used files (such as metadata-files) usually are also held in the server's (SSD) memory, and finally second/backup copies are stored on tapes in a large array storage system. This is transparent (indistinguishable) to the user who might maximally experience slightly different delay times before the visualization on the screen or playback of audio starts. To minimize the risk of loss, the holding is stored automatically in at least six guaranteed identical copies at three different locations: two each in Nijmegen in the Netherlands and in Germany at two large computational centres in Göttingen and Garching. TLA has a guarantee from the Max-Planck-Society that the bit-stream of its collections will be preserved for at least 50 years.

As to the content, the recordings depict natural language use, although the speech events (e.g., the telling of a myth) often have been especially performed for the documentation. Other recordings concern special cultural events or every-day cultural activities, many of which would have taken place identically or similarly (cf. 'the observer's paradox') even without the ongoing documentation. The recordings have then been edited (cut, selected and compiled) by members of the respective teams, with the goal of preserving the original significance and context, cutting out low-quality or irrelevant (parts of) recordings, and those which are inappropriate for archiving because they would, e.g., embarrass or imperil persons depicted or mentioned in the recordings.

So overall, the digital files are as authentic as the technological state of the art allows, and their content is as authentic as the observer's paradox and ethical and quality selection

criteria allow.

5.2 World significance

The significance of the world's linguistic diversity has been repeatedly stressed by UNESCO itself, for instance in the context of the UNESCO Atlas of the World's Languages in danger. The holding here nominated is arguably worldwide the largest collection that tries to capture a representative sample of as much as possible of the world's cultural and linguistic richness.

In terms of depth of individual collections and broadness of the world wide coverage, the proposed holding in its totality is certainly unique and may well be the largest of their kind (depending on the criteria for determining "size"); to our knowledge only the collections at the Endangered Languages Archive (ELAR) at SOAS, London, may have a comparable status. Even so, the holding nominated here is unique in consisting only of large and broad multi-purpose documentations of a generally similar character, focussing on audio and video recordings of natural language use, and for its sound infrastructure (see above, 5.1.). Also, there is little overlap between the two sets of languages covered at ELAR (SOAS) and at TLA's holding here nominated.

Arguably, at this point The Language Archive is the main and best place where researchers can archive their language data with a long-term perspective for their preservation and log-term availability. Other institutions' archives take The Language Archive as a reference model, if they do not themselves apply the technology developed and used at TLA (currently more than a dozen institutions worldwide run archives based on TLA software; SOAS is now in the process of implementing it for the ELAR).

In most cases, the collections in this holding contain documentations made at the last (or even only) opportunity to do so and will be all that will remain of a certain language and culture as the communities change to another dominant language and give up their cultural traditions documented herein. Therefore, the collections are irreplaceable and present also content wise an invaluable cultural heritage for the communities and, in their totality, for mankind.

The collections have already in several cases had great local importance, e.g. supporting language stabilization / strengthening efforts and increasing cultural and linguistic awareness and proud, important factors in stabilizing vulnerable languages. They also serve as a model and inspiration for similar activities in neighbouring speech communities. Finally, they will also be used in expositions, most prominently in the future in the Humboldt-Forum in the reconstructed Berliner Palace where the LandesBibliothek Berlin is setting up a large permanent exposition "World of Languages" where selected materials of this holding will be prominently presented to a large audience (http://www.humboldt-forum.de/en/humboldt-forum/walk-through/). The same will be achieved by ongoing developments of web-portals and apps for mobile devices: The general public is thereby informed about the world's linguistic and cultural diversity. The status of Memory of the World recognized by UNESCO will support The Language Archive in fundraising for such activities.

Also for the history of science, this holding is a landmark, because it is the first major result of the new field of Language Documentation. This field has arguably triggered an important "empirical turn" in linguistics, where linguistic diversity and variability are taken seriously as a basic feature of human culture and cognition, and are not any more down-played or neglected as an epiphenomenon. Linguistic generalizations based on only one or a few languages are not acceptable any more, due to the new focus on the many small languages around the world.

Also, the empirical turn promotes a new scientific posture, where claims about language use and structure have to be verifiable on the basis of easily available basic linguistic data. With this holding, the Language Archive contributed major achievements towards an infrastructure that allows easy and persistent presentation and citation of primary data in linguistic work.

5.3 Comparative criteria:

1 Time

The collections have been made in a moment in global and local history where many cultures and languages in the world are on the verge of extinction – in all continents, small ethnic groups and speaker communities are absorbed in the regional or national society and abandon their traditional language and central cultural features, due to economical, educational and political pressure which leads to, e.g., migration and drastic culture change.

The collections of the holding have specifically been made in order to preserve first-hand knowledge and data on the local languages and traditions before they disappear.

In many cases, the collections also present new facts and discoveries on the languages and cultures they document, but more prominently they are documents of "last of their kind" events.

2 Place

The collections' focus is on small and often unknown or understudied communities, exactly those which are far away from prominent well known places of world history – in the latter, the local cultures and traditions have usually already disappeared. The collections are each of enormous significance for the local culture and communities, and globally as a whole for remembering the heritage of the world's global linguistic and cultural diversity.

3 People

The cultural context of each the collections' creation reflects significant unique aspects of human, in particular linguistic, behaviour. The holding samples the local linguistic and cultural traditions which often developed over many hundreds and even thousands of years. Many collections also feature prominently important key persons in local culture: spiritual and secular leaders, elders, carriers of traditions (some having memorized libraries worth of oral history and literature), persons that possess traditional knowledge that is not acquired by the younger generations, and who usually have high esteem among the community members.

4 Subject and theme

All collections of the holding contain documents of unique local linguistic and cultural traditions; many contain footage of, or information about, important cultural events such as festivals or rituals. In particular, many of the collections contain compilations of the oral literature of a certain people, their myths and other narratives, and the local account of the group's and regional history. Usually, the collections also contain explanations of local costumes and knowledge, for example how to produce important artefacts, on traditional ways of subsistence, or on social, political, religious and spiritual believes and practices.

5 Form and style

Content wise, many of the collections contain unique pieces of art, representing the highest accomplishments of cultural forms such as oratory, poetry, storytelling, singing, dance and similar artistic performances.

The involved research teams have been led by nationally and globally renowned scientists, ensuring very high quality standards with respect to recording, selection and processing of the resources contained in the collections. They serve as a benchmark for similar global and several national language and culture documentation programs, also for the ethical approach of respecting the participating speaker communities as partners in the projects.

The original recordings (primary data) have been enriched by means of annotation (secondary

data) which, if complete, contain a transcription in the original language and translations into (a) nationally and internationally significant language(s), and other annotation of, e.g., an analytical linguistic or commenting / explanatory character. The whole set of resources dedicated to one session (about a speech event, or a cultural event, or a certain topic) has then been described by means of meta-data. Depending on available manpower in the research team and the speech community, the coverage and depth of annotation and meta-data vary from collection to collection, but even the minimally prepared collections already provide a good basis for further research and allow getting important information about the language and culture documented.

A special feature of the holding here nominated is its genuine digital character, so far not often represented in the Memory of the World Register. Also in this technical setting and format, the collections represent a high standard and state-of-the-art in technical infrastructure, data availability and archiving (see above, 5.1.). The technology developed at The Language Archive in order to allow for the proper archiving and making available of the collections has already been worldwide adopted by more than a dozen other regional and even global archives with similar content. For this kind of linguistic and cultural material, it can without exaggeration be said that The Language Archive set a standard that serves as orientation for future developments in archiving technology.

6 Social/ spiritual/ community significance:

The collections have already by now high social significance in the respective communities the languages and cultures of which they depict. Usually the materials have been provided and distributed in an appropriate form (often in digital format at a local cultural centre or as DVDs for public and private exhibition), and in several cases they are being used in schools and language vitalization programs.

It is to be expected that the significance will increase over the years as elders pass away and the collections are the main or only source of information about former cultural features and language use. As we know from many other cases, there may be a cultural turn by the next generations after abandoning the traditional language and cultural features, which harks back and tries to uncover their origin, heritage and identity in a new and broader setting. Under these circumstances, the collections in this holding will be of paramount importance.

Also for the scientific communities (not only linguistics and anthropology, but also several neighbouring fields such as music-ethnology, oral history, ethno-biology, geography, etc.) the collections are more and more being seen and used as data basis for advanced research.

6.0 Contextual information

6.1 Rarity

The materials in this holding are irreplaceable: in most cases, it will not be able to make them again as it takes considerable effort and resources to organize a documentation project while elder speakers pass away and cultural and linguistic practices are being abandoned. The more important it is to preserve these digital resources in a technically sound infrastructure. As detailed above, the digital holding nominated here has an exemplary status with respect to the preservation of important unique and irreplaceable audio-visual content.

Most of the many other recordings from field research among small communities made in the 20th century still remain inaccessible with the researchers or at their institutions. In particular older analogue audio and video recordings, but also not well curated and archived digital material is under an enormous risk of being lost forever – due to deteriorating data carriers, obsolete technology or outdated file formats. Some 80% of valuable recordings are in danger of being lost, as was reported in 2004 by Dietrich Schüller, Chair of the UNESCO Memory of the World Sub-Committee on Technology. Despite many digitization efforts, not much has changed globally since then.

6.2 Integrity

See comments on (especially technical aspects of) "authenticity", 5.1 above. Generally, the holding is technically integer and content wise faithful to the languages and cultures therein documented.