

UNESCO Memory of the World: Selected Data Collections of the World's Language Diversity at The Language Archive

Paul Trilsbeek

The Language Archive

Max Planck Institute for Psycholinguistics



DOBES

DOKUMENTATION BEDROHTER SPRACHEN
DOCUMENTATION OF ENDANGERED LANGUAGES



United Nations
Educational, Scientific and
Cultural Organization



- Selected data collections
- of the world's language diversity
- at the Language Archive
- Inscribed on the Register in 2015
- [Memory of the World](#)

The Language Archive

- Digital archive of language materials based at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
- Archive exists since the late 90's, initially archiving language materials from MPI's own field researchers and language acquisition researchers
- Became the central archive for the DOBES endangered languages documentation programme, funded by the Volkswagen Foundation in 2000

DOBES Programme

- Large funding programme by the Volkswagen Foundation for the documentation of endangered languages (DOKumentation BEdrohter Sprachen)
- Started in 2000, will finish in 2019 (last new projects started in 2012)
- 67 language documentation projects funded, several documented more than one language

The Language Archive Collections

- Holds about 370 collections covering more than 200 different languages
- More than 17000 hours of audio and video recordings
- About 90 TB of data
- Indigenous/Endangered languages
- Language Acquisition studies
- Sign Language collections
- Experiment data collections

Long-term Preservation

- Digitisation of analogue media (tape, film, etc.) before they deteriorate
- Digital preservation:
 - preservation of bit-streams -> data replication, timely migration of storage media
 - Preservation of content -> file format migration before they become obsolete
- UNESCO Recommendation Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form (2015)

Long-term Preservation at TLA

- Digital archive only: anything analogue will be digitised first
- 6 copies of each file in 5 different buildings in 3 different geographical locations in the Netherlands and Germany
- Copies stored on a mixture of storage media that allow for automated migration (Hard disk arrays, Solid State Drives, Data tape)
- Monitoring of file integrity
- Supported file formats chosen with long-term preservation characteristics in mind
- Monitoring of file format obsolescence

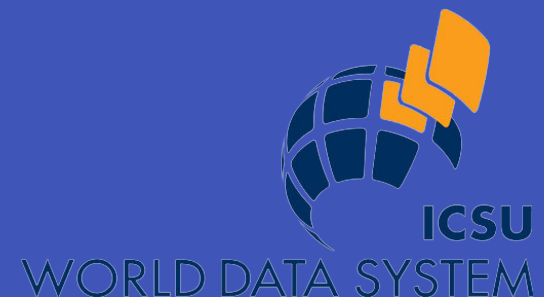
DELAMAN Project Lost & Found

- Digital Endangered Languages and Musics Archives Network
- 12 full and 2 associate member archives
- Project Lost & Found: identifying valuable at risk collections and looking for digitisation and archiving possibilities among the DELAMAN members

-> www.delaman.org

TLA recognition

- Assessed according to Data Seal of Approval standard for trusted digital repositories
- Regular member of the ICSU World Data System
- Certified CLARIN B-type Centre



UNESCO Memory of the World

- October 2015: Selected collections of TLA inscribed to the UNESCO Memory of the World Register
- 64 collections, containing materials from 102 different languages
- Approx. 3000 hours of video, 2000 hours of audio, 43000 photographs, 17000 written documents
- Great recognition of the value of these collections for the world, as well as of the work that TLA has done in preserving them

TLA Memory of the World Collections



United Nations
Educational, Scientific and
Cultural Organization



- Selected data collections
- of the world's language diversity
- at the Language Archive
- Inscribed on the Register in 2015
- *Memory of the World*

Example 1: Fish Trap Making

- Language: Bakola, Gyeli dialect
- Location: Lolodorf region, Republic of Cameroon
- Number of speakers: 2200-5000
- Vitality: endangered

- Speaker: Lène
- Recorded by: Nadine Borchardt and Christopher Lorenz

Example 2: Flower Sand Drawing

- Language: Dal kalaen
- Location: Wuro village, West Ambrym, Vanuatu
- Number of speakers: approx. 1000
- Vitality: threatened. Children still acquiring it as a first language

- Speaker: Diana
- Recorded by: Soraya Hosni

TLA Memory of the World Languages

≠Akhoe Hai//om	Daakaka	Kosarek-Yale	Papuan Malay	Tsengel Tuvan
Aché	Dal kalaen	Kri	Resigaro Arawak	Tsova-Tush
Afro-Ecuadorian Spanish	Déline	Kuikuro	Sanzhi	Tuvan
Akie	Eipo	Laal	Savosavo	Tzeltal
Akkala Saami	Enets	Lao	Semai	Udi
Alu Kurumba	Even	Lokono	Shiri	Vilela
Aweti	Gela	Maniq	Singpho	Waimaqa
Bainounk Gubëeher	Guugu Yimithirr	Marquesan	Skolt Saami	
Bainounk Gujaher	Ilgar/Garig	Marrku	Sochiápam Chinantec	
Bakola	Inuktitut	Minderico	Sri Lanka Malay	
Baure	Irohito	Mocovi	Svan	Wichi
Beaver	Isubu	Movima	Taa	Wichita
Boor	Iwaidja	Mpur	Tai	Witoto proper
Bora	Jahai	Mullu Kurumba	Tangsa	Wooi
Boua	Jenu Kurumba	Murato	Tapiete	Wovia
Cashinahua	Kata Kolok	Ndam	Ter Saami	Yali
Cha'palaa	Katang	Ngeq	Tima	Yauyos Quechua
Chatino	Khinalug	Nonuya	Tofa	Yélî Dnye
Chipaya	Khmer	Nsong	Totoli	Yucatec
Cholanaika	Kildin Saami	Ocaina	Tozhu	Yurakaré
Chontal	Kilivila	Oyda	Tsafiki	Zapotec
	Kómnzó			

Further information: tla.mpi.nl