



Das Spracharchiv des MPI-PL Nimwegen
Hintergrund und Aktivitäten

UKP, U. Darmstadt, 21. Juni 2012

SEBASTIAN DRUDE
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Das Spracharchiv (TLA) @ MPI-PL

1. Hintergrund: Sprachenvielfalt, Sprachbedrohung, Sprachdokumentation
2. Datenarchiv und Digitalisierung
3. Software & Infrastruktur
4. Projekte, Netzwerke und Zusammenarbeit
5. Fazit: Das Spracharchiv am MPI-PL

Das Spracharchiv (TLA) @ MPI-PL

1. Hintergrund: Sprachenvielfalt, Sprachbedrohung, Sprachdokumentation
2. Datenarchiv und Digitalisierung
3. Software & Infrastruktur
4. Projekte, Netzwerke und Zusammenarbeit
5. Fazit: Das Spracharchiv am MPI-PL

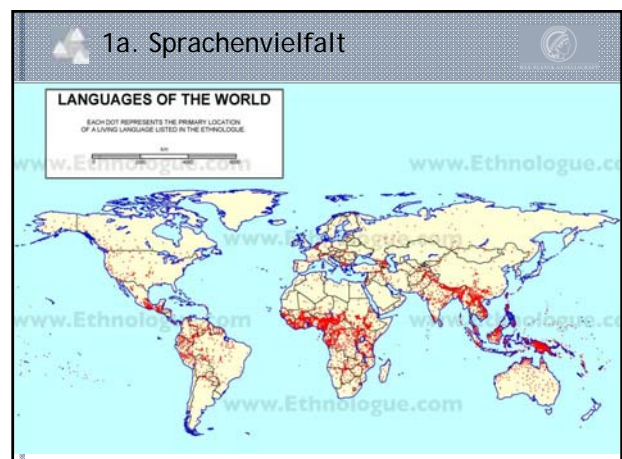
1a. Sprachenvielfalt

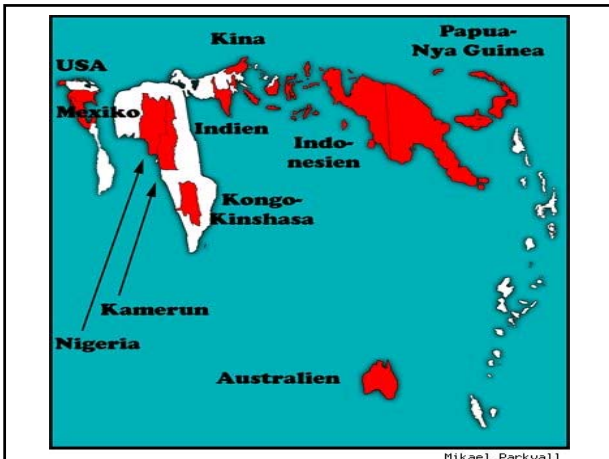
- Es gibt weltweit 5.000–7.000 Sprachen
- Verschiedene Zählungen / Schätzungen
- Exakte Zahlen sind schwierig
 - fehlendes Wissen
 - Sprache-Dialekt-Problem
- Die Sprachen gehören zu **Sprachfamilien**:
Es gibt 90–150 Sprachfamilien, dazu mindestens 36 isolierte Sprachen, 82 Kreolsprachen und 119 Gebärdensprachen

1a. Sprachenvielfalt

Sprachen sind geographisch sehr ungleich verteilt

Area	Living languages		Number of speakers			
	Count	Percent	Count	Percent	Mean	Median
Africa	2,092	30.3	675,887,158	11.8	323,082	25,391
Americas	1,002	14.5	47,559,381	0.8	47,464	2,000
Asia	2,269	32.8	3,489,897,147	61.0	1,538,077	10,171
Europe	239	3.5	1,504,393,183	26.3	6,294,532	220,000
Pacific	1,310	19.0	6,124,341	0.1	4,675	800
Totals	6,912	100.0	5,723,861,210	100.0	828,105	7,000

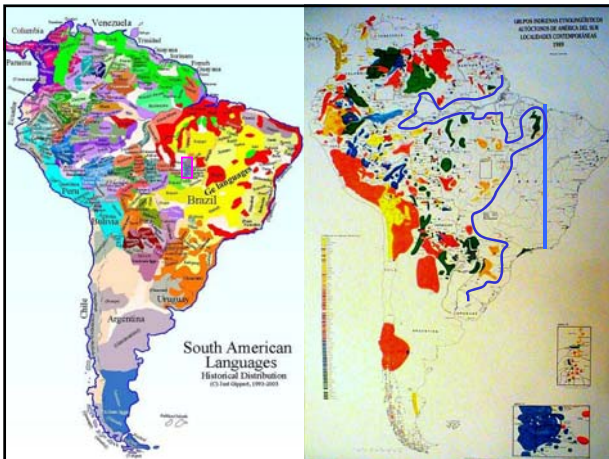




1a+b. Sprachenvielfalt, -bedrohung

Sprachen sind demographisch sehr verschieden

Population range	Living languages			Number of speakers		
	Count	Percent	Cumulative	Count	Percent	Cumulative
100,000,000 to 999,999,999	8	0.1	0.1%	2,301,423,372	40.20753	40.20753%
10,000,000 to 99,999,999	75	1.1	1.2%	2,246,597,929	39.24969	79.45723%
1,000,000 to 9,999,999	264	3.8	5.0%	825,681,046	14.42525	93.88247%
100,000 to 999,999	892	12.9	17.9%	283,651,418	4.95560	98.83807%
10,000 to 99,999	1,779	25.7	43.7%	58,442,338	1.02103	99.85910%
1,000 to 9,999	1,967	28.5	72.1%	7,594,224	0.13268	99.99177%
100 to 999	1,071	15.5	87.6%	457,022	0.00798	99.99976%
10 to 99	344	5.0	92.6%	13,163	0.00023	99.99999%
1 to 9	204	3.0	95.5%	698	0.00001	100.00000%
Unknown	308	4.5	100.0%			< 0.2%
Totals	6,912	100.0		5,723,861,210	100.00000	



1b. Sprachbedrohung

Weltweit ist die Sprachenvielfalt bedroht:

- Schätzung: 50% sind bereits stark bedroht und werden bis 2100 verschwunden sein
- Weitere 40% können 2100 ihrerseits 'moribund' sein (nur erwachsene Sprecher)
- Das heißt, 2150 gibt es voraussichtlich nur noch 400-900 Sprachen in der Welt

1b. Sprachbedrohung

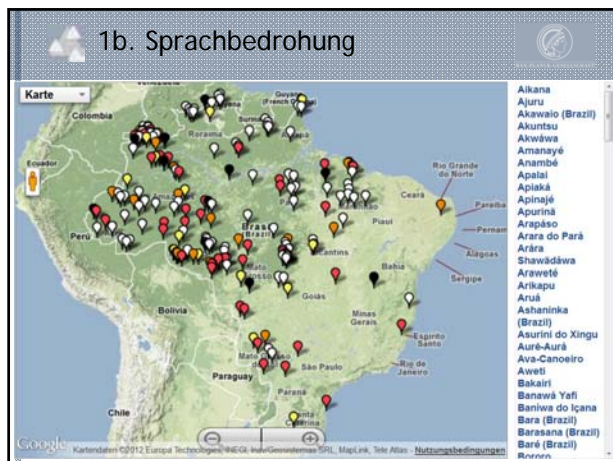
UNESCO (2003) "Language Vitality and Endangerment" – Die wichtigsten (von 9) Kriterien:

1. Absolute Sprecherzahlen
2. Anteil der Sprecher an der allgem. Bevölkerung
3. Intergenerationelle Weitergabe
4. (Änderungen in der) Domänenverteilung
5. Reaktionen auf neue Domänen (neue Media...)
6. Lehrmaterial, Alphabetisierung

1b. Sprachbedrohung

UNESCO Atlas of Endangered Languages (Moseley, Ed.)

Degree of endangerment	Intergenerational Language Transmission
safe	language is spoken by all generations; intergenerational transmission is uninterrupted >> not included in the Atlas
vulnerable	most children speak the language, but it may be restricted to certain domains (e.g., home)
definitely endangered	children no longer learn the language as mother tongue in the home
severely endangered	language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
critically endangered	the youngest speakers are grandparents and older, and they speak the language partially and infrequently
extinct	there are no speakers left >> included in the Atlas if presumably extinct since the 1950s



1c: Sprachdokumentation

Ziele:

- Abstrakt: "Das intellektuelle Erbe der Menschheit aufzeichnen / festhalten"
- Die traditionellen Überlieferungen für die Sprecher und ihre Erben bewahren
- Daten für empirische linguistische Untersuchungen liefern (*Corpora*)
- Daten für benachbarte Disziplinen liefern (Anthropologie, Geschichte, Botanik...)
- Materialien für Spracherhaltsmaßnahmen

1c: Sprachdokumentation

Zukünftige Anwendungen sind unvorhersehbar, vielleicht neuartig

Fazit:
Sprachdokumentationen sollten im Idealfall einer Vielzahl von unterschiedlichen Zwecken dienen ("*multipurpose*")

1c: Sprachdokumentation

Zeitliche Entwicklung:

- 1989-92 : Interesse für und Sorge um „*Endangered Languages*“ beginnt
- 1993: Sommerschule in Köln (GfBS & VWS)
- 1999: erste Ausschreibung für DOBES (VWS)
- 2000: AILLA @ Univ. Texas; E-MELD
- 2002: HRELDP @ SOAS / London
- 2003: UNESCO gibt Dokument heraus: "Language Vitality and Endangerment"
- 2004 ss.: DELAMAN, regionale Archive und Netzwerke, CLARIN

1c: Sprachdokumentation

Tätigkeiten und Methoden:

- Feldforschung, Aufnahme von Primärdaten
- Annotation und Analyse: Sekundärdaten
- Organisation und Archivierung
- Sprachplanung, -revitalisierung etc.
- (Auswertung, Analyse, wiss. Publikationen)

1c: Sprachdokumentation

Daten
Daten sind immer Daten FÜR etwas, zumindest VON etwas – üblicherweise **eine systematische Repräsentation von** realweltlichen Zuständen und Ereignissen
In der Linguistik sind **Primärdaten** meist eine direkte Wiedergabe oder das Resultat eines **kommunikativen (Sprech-) Ereignisses**, zum Beispiel ein **geschriebener Text** oder, insbesondere, eine **Audio/Video-Aufnahme** eines Sprechereignisses

1c: Sprachdokumentation

Annotation & Sekundärdaten:
 Die **Annotation (A)** einer Aufnahme eines Sprechereignisses S ist eine **schriftliche / symbolische Repräsentation** von Eigenschaften von S
Globale A. bezieht sich auf S als ganzes, inkl. der Umstände etc. (Teil der **Metadaten**)
Temporale A. repräsentiert Eigenschaften von einzelnen Äußerungen von S oder deren Teile

1c: Sprachdokumentation

Annotation & Sekundärdaten, Metadaten:
 Nachhaltige Annotation – Standards und Empfehlungen:

- Offene (und freie) Formate
- Nicht-binäre Formate (textbasierte Formate)
- **XML** (mit geeigneten Dokumentdefinitionen)
- Kodierung: **UNICODE**
- Abkürzungen, Konventionen explizit machen

1c: Sprachdokumentation

Annotation & Sekundärdaten:

- Es gibt verschiedenste Schemata für Ebenen
- Minimum: **Transcription + Übersetzung**
- **Typologische Glossierung:**
Interlinear Morphemic Translations :
 (a) Transkription
 (b) einzelne Morphe
 (c) Glosse(n) für jedes Morph
 (d) [idiomatische] Übersetzung
- Andere Ebenen: lautbezogen, grammatisch, ...

1c: Sprachdokumentation

The screenshot shows a detailed linguistic analysis of the sentence "E: it was like this that our grandfathers lived." The interface includes multiple layers of annotation: phonetic transcription (e.g., 'j a t a ʰ t s u j a t a : a z o a m u j z a n e k o z o k w a w u t n e ʰ a ʔ e'), orthographic transcription, word categories (e.g., DEM, PP, SUB, VI, 3.Past), and constituent structure (e.g., 1.2 Phr, 1.2.1, 4NF, 5VF, 6Ph, 4.5VPr, 4.5.6Vgr, 1.2.1.4.5.6.7). A comment at the bottom explains that the first part of the sentence is a subordinate clause.

1c: Sprachdokumentation

The screenshot shows the ELAN software interface. At the top, there is a video window showing two people in a conversation. Below the video, there are several tracks for audio analysis, including a pitch contour plot and a transcription track. The transcription track shows the text of the conversation with time markers and alignment with the video frames.

1c: Sprachdokumentation

Fortschritte und Stand:

- Die grundsätzlichen Ziele und Methoden sind etabliert: daß/wie man Digitale Archive mit Primärdaten und Annotation aufbaut
- Relevante Software wurde und wird entwickelt (ELAN, IMDI, LAT)
- Rechtliche und ethische Fragen werden bedacht und meist beantwortet
- Nationale u. regionale Archive entstehen
- Bewußtsein für Sprachvielfalt gestärkt
- Aber: Geldgebende Programme enden bald

Das Spracharchiv (TLA) @ MPI-PL

1. Hintergrund: Sprachenvielfalt, Sprachbedrohung, Sprachdokumentation
2. Datenarchiv und Digitalisierung
3. Software & Infrastruktur
4. Projekte, Netzwerke und Zusammenarbeit
5. Fazit: Das Spracharchiv am MPI-PL

2. Datenarchiv und Digitalisierung

Gegenwärtiger Status des Archivs:

- ca. 70 TeraByte gut beschriebener Ressourcen
- ca. 22.000 h digitaler Audio- / Videoaufnahmen
- ca. 75.000 "Sessions" (Metadatensätze)
- ca. 5 Mio. annotierte Segmente
- ca. 50 Lexica
- Wir nehmen jederzeit Legacy-Datensätze zur (Digitalisierung und) Archivierung an

2. Datenarchiv: Datentypen

"Sessions" und Metadaten

SESSION

Metadaten
(beschreiben das Ereignis und die dazugehörigen Daten)

Videoaufnahme

Audioaufnahme

Annotation

PRIMÄRDATEN

SEKUNDÄRDATEN

Transcription: Ortografisch / Phonolog. ...

Wörtliche / Idiomatiche Übersetzung ...

(linguistischer / ethnograph. Kommentar...)

(Morphem-Glossen...)

...

2. Datenarchiv: Status

- Ressourcen zu ca. 200 Sprachen
- ca. 65 DOBES-Projekte dokumentieren über 85 bedrohte Sprachen
- Spracherwerbs-, Gesprächs-, multimodale, mehrsprachige, ethnologische etc. Daten, nun auch Brain-Imaging, genetische Daten
- Das Archiv wächst derzeit jährlich um ca. 5-10TB

2. Datenarchiv: IMDI-Browser

The screenshot shows the IMDI browser interface. On the left is a navigation tree with categories like 'From Language and Culture', 'Linguistic Data', 'Monological', and 'Biographies'. The main content area displays details for a session named '026_autobiogr'. The title is '026 tells the story of his life' and the date is '2002-09-20'. The description states it is an autobiographical recording by 026, with audio and video files. The location is 'Project Aweti' and the content is 'basic annotation Port. only'. The actors section mentions 'SD asks the consultant, 026, to tell his autobiography. 017 translates between SD and 026.'

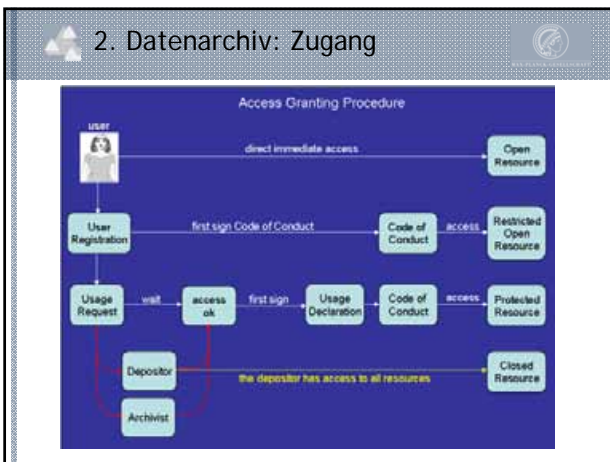
The screenshot shows a file explorer view of the archive's directory structure. It is organized hierarchically under 'From Language and Culture'. Key folders include 'Cultural Data', 'Artifacts+Housing', and a large 'Aweti' folder. Under 'Artifacts+Housing', there are subfolders for '026_house01' through '026_house13'. The 'Aweti' folder contains numerous subfolders for different linguistic and cultural aspects, such as 'estrelka', 'hammock', 'house_006', 'house_118', 'house_132', 'house_burning', 'making_basket', 'stool', 'urumum04', 'wak_scoops', 'Culture Contact', 'dist_present', 'guest_house01' through '04', 'ling_worh04', 'mko_003_sham', 'mko_hammock', 'neto_waura', 'portb_jeonadco', 'pres-ent-05', 'radio_set', 'soccer2', 'soccer_01B', 'solar_jigs', 'wordcup_TV', and 'Everyday Culture'.

2. Datenarchiv: Annex

The screenshot shows the ANNEX software interface. It features a central video display window showing a person speaking. To the right, there is a 'Media Information' panel with fields for Resource, Media file, Elapsed time, Selected chunk, Start time, End time, and Text. Below the video is a 'Waveform and Timeline' section with two channels (Left channel and Right channel) and a 'Midi Data Frame' section. The interface includes various playback controls and a 'Text' panel on the left.

2. Datenarchiv: Perspektive

- Für DOBES Daten werden derzeit sechs Kopien automatisch an drei Orten erstellt
- Garantie der MPG für 50 J. Bit-stream-Erhalt
- Digitale Datenzentren: je öfter das Material aufgerufen und benutzt wird, desto besser
- ⇒ Wichtiges Ziel: einfacher und nützlicher Zugang
- Aufgabe für Techniker UND Wissenschaftler
- Ein wichtiger Aspekt einer fruchtbaren Beziehung zwischen Forscher und Zentrum ist **Vertrauen**



- ### Das Spracharchiv (TLA) @ MPI-PL
1. Hintergrund: Sprachenvielfalt, Sprachbedrohung, Sprachdokumentation
 2. Datenarchiv und Digitalisierung
 3. Software & Infrastruktur
 4. Projekte, Netzwerke und Zusammenarbeit
 5. Fazit: Das Spracharchiv am MPI-PL

3. Software

- Unterstützung für den vollen "Lebenszyklus"
- Neues Metadatenformat: CMDI (flexibel, Komponenten können ausgesucht werden)
- ARBIL unterstützt CMDI
- AVATech: Module für automatische A/V Erkennung
- Nun Lexan: (semi-)automatische Annotation

The diagram shows the software lifecycle: 'preparation' (ELAN / LEXUS Annotation + Lexicon), 'integration' (IMDI / CMDI / ARBIL Data Organization Metadata Description, Archival Infrastructure, LAMUS Data Uploading and Management Access Management), 'Data Archiving and Copying', and 'utilization' (IMDI / CMDI / GIS / VLO Metadata Browsing & Searching, ANNEX / LEXUS / IMEX / TROVA Complex Access via Web, RELcat / ISOut Ontology management framework, VICOS Semantic Access and Enrichment).

3. Software: ELAN annotation tool

The screenshot shows the ELAN annotation tool interface. It includes a 'video player' at the top, 'controls' below it, a 'crosshair' for navigation, a 'waveform viewer' showing audio waveforms, 'annotations' for marking specific points in time, and a 'timeline viewer' at the bottom. A red circle highlights the 'tiers' section of the timeline viewer.

3. Software: ELAN annotation tool

The screenshot shows the ELAN software interface. At the top, there's a menu bar with options like File, Edit, Annotation, Tools, View, Options, Window, Help. Below the menu is a toolbar with icons for various functions. The main area is divided into several panes: a video player on the left showing a person, a list of annotations in the middle, and a timeline at the bottom with colored bars representing different annotation layers.

3. Software: Lexus (Online-Lexicon)

- Erstellung neuer Lexika, Import von Lexica in anderen Formaten (Toolbox, XML, Chat)
- Konfigurierbare Ansichten: Wortliste und einzelner Eintrag

The screenshot displays the Lexus Online-Lexicon web interface. It features a search bar at the top, a list of words on the left, and a detailed view of a selected word on the right. The word view includes a definition, examples, and a small image. The interface is clean and organized, with clear navigation options.

3. Software: Lexus (Multimedia)

- Verknüpfung von Multimedia-Elementen mit Einträgen

The screenshot shows the Lexus Multimedia interface. It displays a word entry with a list of multimedia elements (audio, video, image) and links to archived files. A legend on the right side identifies the icons for Audio, Video, Image, and Link to archived files.

3. Software: Vicos (Konzepträume)

The screenshot shows the Vicos software interface, which is a concept space tool. It displays a network diagram of concepts connected by lines. The nodes in the network are labeled with words and their corresponding URLs, such as 'http://lex.lug.uni-erlangen.de/lexicon/lexicon/172_1624_131119729258091305C0006L_8%'. The interface includes a menu bar and a toolbar.

3. Software: Arbil (Metadatenedit.)

The screenshot shows the Arbil Metadatenedit interface. It features a 'Word list' view on the left and a 'Lexicon: Word list' view on the right. The word list view includes a search bar and a list of words. The Lexicon view shows a detailed entry for a word, including its definition, examples, and a small image. The interface is designed for editing and managing lexical data.

3. Software: Arbil (Metadatenedit.)

The screenshot shows the Arbil Metadatenedit interface, focusing on the metadata editing window. It displays a list of metadata fields and their values, allowing users to edit and manage the metadata for their lexical entries. The interface is complex and detailed, with many options for customization.

3. Software / Infrastructure: Trova

Simple search

The screenshot shows the Trova search interface. At the top, there are tabs for 'Simple', 'Single Layer', and 'Multiple Layer'. The 'Simple' tab is selected. Below the tabs, there is a search bar with the text 'grater' and a 'Find' button. The search results are displayed in a table with columns for 'File', 'Size', and 'Progress'. The search results show a list of files with their names and sizes.

3. Software / Infrastructure: Trova

Show Frequency View

The screenshot shows the Trova search interface with the 'Show Frequency View' selected. This view displays a list of files with their names and sizes, grouped by frequency. The search results show a list of files with their names and sizes.

3. Software / Infrastructure: Trova

Regular expression:
 $[^n]g$$ finds all ending 'g', but not 'ng'

The screenshot shows the Trova search interface with a regular expression search. The search bar contains the regular expression $[^n]g$$. The search results are displayed in a table with columns for 'File', 'Size', and 'Progress'. The search results show a list of files with their names and sizes.

3. Software / Infrastructure: Trova

Advanced search options

The screenshot shows the Trova search interface with advanced search options. The search bar contains the regular expression $[^n]g$$. Below the search bar, there are several options for refining the search, including 'Model', 'Initial Duration', and 'Final Duration'. The search results are displayed in a table with columns for 'File', 'Size', and 'Progress'.

3. Infrastructure: Copies

The diagram shows a network of computer centers and servers. A central hub is connected to several computer centers, each with its own set of servers. The diagram illustrates the distribution of data across multiple locations.

- Stabiles, robustes, kohärentes, organisiertes Online-Archiv mit 80 TB Ressourcen für eScience
- Alle mit Metadaten beschrieben und mit PIDs versehen
- 4 volle dynamische Kopien bei Comp.Cent. mit 50 J. Garantie
- Zusätzlich 11 regionale Archive (Zahl wächst weiterhin)
- Offener Deposit-service

3. Infrastruktur: Replikation

The diagram shows data replication between different zones and servers. It illustrates the flow of data from a central source to multiple destinations, ensuring redundancy and availability.

Arbeiten an sicherer Replikation mit PIDs und Checksums als Basis

Vor und nach jeder Operation werden die Checksummen geprüft

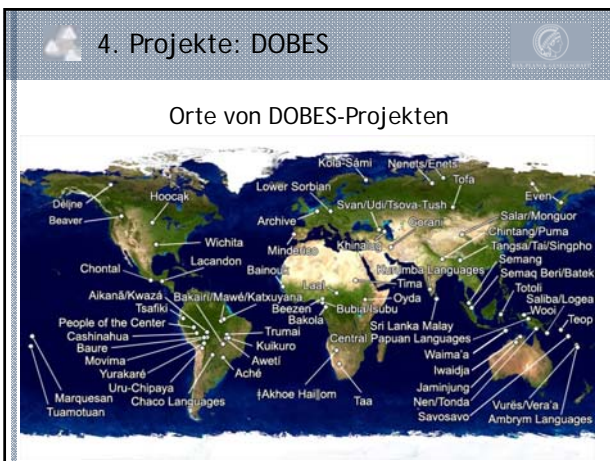
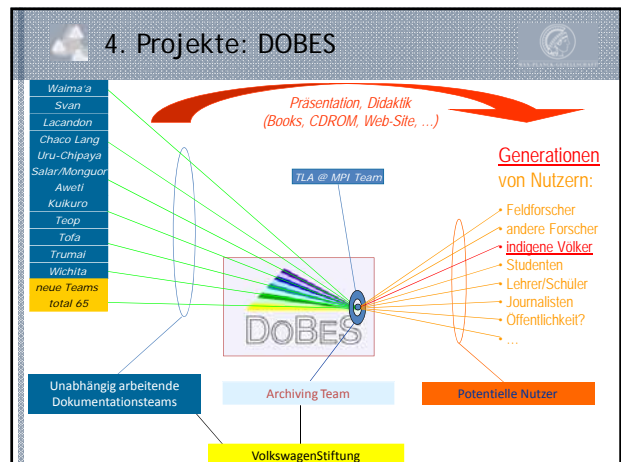
3. Software / Infrastruktur: Status

	state	
LAMUS	ausgereift	viele Archiv-Föderation-Aktivitäten
AMS	ausgereift	keine Aktivitäten
IMDI	veraltet	wird nach und nach ersetzt
ARBIL	ausgereift	wird ausgebaut
CMDI	in Entwickl.	wird IMDI ersetzen
ELAN	ausgereift	durch A/V-Erkennen und Lexan erweitert
ANNEX	ausgereift	Erweiterungen möglich
IMEX	ausgereift	keine Entwicklungen
TROVA	ausgereift	soll in AVATech erweitert werden
LEXUS	überarbeitet	komplette Überarbeitung nun abgeschlossen, testing
VICOS	überarbeitet	verbesserte Integration schreitet fort
ISocat	ausgereift	einige zusätzliche Entwicklungen
Lexan	begonnen	LEXUS-ELAN/ANNEX-Integration wird entwickelt
ADDIT	gestoppt	nicht mehr unterstützt

- ### Das Spracharchiv (TLA) @ MPI-PL
1. Hintergrund: Sprachenvielfalt, Sprachbedrohung, Sprachdokumentation
 2. Datenarchiv und Digitalisierung
 3. Software & Infrastruktur
 4. Projekte, Netzwerke und Zusammenarbeit
 5. Fazit: Das Spracharchiv am MPI-PL

- ### 4. Projekte
- ISLE, MUMIS, ECHO, INTERA, DAM-LR, **DoBeS**, CGN, **CLARIN EU, CLARIN NL, CLARIN D**, CLARA, Inter, HARVE, AVATech, REPLIX, RELISH
 - EUDAT, DASISH, INNET, iCORDI
 - Standardisierungsarbeit in ISO 3007/ISO 15926
 - ISO DCR, MPI-PL ist Registration
 - Standards: 12620 DCR, 24618 PI
- Why Infrastructure projects?

 - Data is the treasure for empirical research
 - Allow our researchers access to the best data & tools
 - Fair exchange of efforts
 - Important to embed our archive in the wider world



4. Projekte: DOBES: Zugang

Attraktive Community-Portals

4. Projekte: INNET

Initiative "Regionale Archive": Zusammenarbeit TLA/MPI-PL mit anderen Organisationen, die an Bed. Spr. Datenarchivierung interessiert sind
 Auch größere Institutionen verwenden nun TLA (LAT) Software

- Unterstützung der lokalen Produktion und Archivierung von Sprachdaten
- Ziel: Austausch von Daten und Ideen

Das Spracharchiv (TLA) @ MPI-PL

1. Hintergrund: Sprachenvielfalt, Sprachbedrohung, Sprachdokumentation
2. Datenarchiv und Digitalisierung
3. Software & Infrastruktur
4. Projekte, Netzwerke und Zusammenarbeit
5. Fazit: Das Spracharchiv am MPI-PL

5. Fazit: Die Max-Planck-Gesellschaft

- Über 80 "MPI's": Institute für Grundlagenforschung
- Alle wiss. Disziplinen:
 - Naturwissenschaften
 - Lebenswissenschaften
 - Sozialwissenschaften
 - Geisteswissenschaften
 - Recht, Mathematik, Technik, ...
- Fast alle in Deutschland, wenige im Ausland (I, NL, USA)
- MPI für Psycholinguistik in Nijmegen, Niederlande

5. Fazit: Das Spracharchiv: Geschichte

- Ursprünglich "Technical Group" am MPI-PL
- 1990er: Herausforderung der Archivierung von digitalen Daten erkannt
- Sprachdokumentation: DoBeS ca.2000–ca.2016
- Seit 2000: Entwicklung von Software: "Language Archiving Technology"
- Notwendigkeit der Nachhaltigkeit des Archivs und der Aktivitäten (Softwareentwicklung, Infrastruktur etc.)
- September 2011: Neue Einheit eröffnet

5. Fazit: Das Spracharchiv: Finanzierung

Grundfinanzierung in binationaler Kooperation:

• MPI Nijmegen (W. Klein)	2
• Max Planck Society	2,5
• Berlin Brandenburg. Akademie d. Wiss.	2
• Königl.-Niederländ. Akademie der Wiss.	2

Externe Projekte finanzieren weiterhin den Großteil der Arbeit und 28 Mitarbeiter

5. Fazit: Das Spracharchiv : Perspektive

- Grundförderung sichert ca. 8 feste Positionen
- Diese Förderung geht über 5 Jahre
- Evaluation der Qualität nach 3-4 Jahren
- Zweck: Angestrebt wird klar eine Verstetigung zur Nachhaltigkeit der Daten und Software

 5. Fazit: Das Spracharchiv: Menschen 



• Wolfgang Klein	Wiss. Direktor	3 j.
• Peter Wittenburg	Head of Unit	2 j.
• Daan Broeder	Deputy Head	
• Sebastian Drude	Scientific Coordinator	

Core Team and responsibilities:



• Peter Wittenburg	management, projects
• Daan Broeder	technology, archive software
• Sebastian Drude	bridge to users, project mng.
• Paul Trilsbeek	archive, digitization
• Han Sloetjes	annotation software
• Menzo Windhouwer	knowledge software

 5. Fazit: Das Spracharchiv: Menschen 

Archivsoftware u. Ä. (LAMUS, AMS, IMDI, CMDI, ARBIL, ...):	Archiv, Digitalisierung, Softwaretesting, Manuals:
• <u>Daan Broeder</u>	• <u>Paul Trilsbeek</u>
• Peter Withers	• Nick Wood
• Dieter van Uytvanck	• Alexander König
• Twan Goosen	• Florian Wittenburg
• Guilherme da Silva	• Chris Haskett
• Jean-Charles Ferrieres	• Vlado Plaga
• Micha Hulsbosch	• Peter Nijland
• Mariano Gardellini (LAMUS)	• Jeroen Geerts (Mai '12)
• Patrick Duin (LAMUS, VLO)	• ca. 4 stud. Hilfskräfte
• Willem Elbers	
• Gerd v.d. Plas	
• Sander Meijers (Jun. '12)	

 5. Fazit: Das Spracharchiv: Menschen 

Annotationssoftware (ELAN, ANNEX, IMEX, TROVA, ...):	Knowledge-software (LEXUS, VICOS, ISOcat):
• <u>Han Sloetjes</u>	• <u>Menzo Windhouwer</u>
• Aarthi Somasundaram	• André R. A. Moreira
• Lari Lampen	• Claus Zinn (VICOS)
• Thomas Koller (ANNEX)	• Jacquelijin Ringersma (LEXUS/VICOS, Archiv)
• Kees Jan van de Looij (Jun '12)	• Marc Kemp-Snijders (LEXUS)
AV-detectors	• Huib Verweij (LEXUS)
• <u>Przemek Lenkiewicz</u>	• Herman Stehouwer
• Binyam G. Gebre	• Shakila Shayan (Mai '12)
• Anna Jaworska	• Olha Shkaravska (Juli '12)
• Eric Auer	

 5. Fazit: Das Spracharchiv: Mission 

- Digitalisierung und Archivierung wertvoller Sprachressourcen von weltweiter Forschung
- Zugang zu archivierten Materialien unter Beachtung rechtlicher und ethischer Beschränkungen
- Entwicklung von Werkzeugen, Diensten und Infrastruktur zum freien Gebrauch in der Forschung
- Weltweit Einrichtung von Archiven um Sprachdokumentation und -revitalisierung zu fördern
- Einsatz für internationale Standards und Entwicklung von digitalen Forschungs-Infrastrukturen
- Ausrichten von und Teilnahme an Ausbildungs- und Trainingskursen, Hilfe und Support

Das Spracharchiv des MPI-PL Nimwegen Hintergrund und Aktivitäten

UKP, U. Darmstadt, 21. Juni 2012

SEBASTIAN DRUDE
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands