Eureka! User friendly access to the MPI linguistic data archive

Alexander Koenig
Jacquelijn Ringersma
Claus Zinn

Max Planck Institute for Psycholinguistics
Misconception about archiving

Your stuff is buried here and gone forever
• only 10 years since we started
• 300,000 audio, video & text resources
• 100,000 metadata descriptions
• 16.5 TB of data

...and still growing daily
MPI researchers' data
DoBeS projects
Corpora from other research institutions
  • Corpus Nederlandse Gebarentaal
  • European Science Foundation's Second Language Acquisition by Adult Immigrants
  • Dutch Bilingual Database
Diverse Sources - One Structure

• all data described using IMDI metadata format
• central organisational concept: the session resource bundle
  – obligatory metadata
  – a set of annotations & transcriptions
  – a set of media data (video, audio, photo)
• every resource and metadata description has a URID (http://www.handle.net/)
• all data in principle available online
  – http://corpus1.mpi.nl
• complex access management system regulates who can access what
Accessing the Data

- TROVA
- IMDI Browser
- Metadata Search
- Google Earth Overlay
- Facetted Browsing
- LEXUS
- Web Portals
- ViCoS
Accessing the Data

Linguistic Researchers / Depositors

TROVA

Metadata Search

IMDI Browser

Researchers from other Domains

Web Portals

Google Earth Overlay

Speech Communities

ViCoS

Facetted Browsing

LEXUS

The General Public
Accessing the Data

The Traditional Method

- The IMDI Browser
• data is structured by researchers
• can be ordered by geographic location or thematic subject or combination of these
• Metadata is directly visible from the IMDI browser
Resources can be viewed using ANNEX
More than just Browsing

Metadata and Content Search
• Metadata is always open and can be searched
If you have access to the resources, you can also search the annotations.
Accessing the Data

The Narrow-down Method
—
Facetted Browsing
• facets are taken from IMDI metadata
• users can browse through data
  – user selects facet (i.e. genre: conversation)
  – only matching sessions are shown
  – user selects another facet (i.e. country: Brazil)
  – an even smaller subset is shown
The Narrow-down Method – Facetted Browsing

IMDI Archive (Demonstrator) (Flamenco)

CORPUS
- alla (1794)
- andal (231)
- bas (7427)
- bibl (1178)
- cong (92029)
- cors / Rom (772)
- dhol (2122)
- dobos (17498)
- edo (245)
- endangered Languages (12941)
- est (2854)
- gtu (34)
- more...

CONTINENT
- Africa (2762)
- Asia (13420)
- Australia (4822)
- Europe (124516)
- Middle America (1737)

COUNTRY
- Argentina (816)
- Australia (4839)
- Belgium (23444)
- Bolivia (5013)
- Botswana (504)
- Brazil (3435)
- Canada (3427)

LANGUAGE
- akk (348)
- lu (2)
- dhil (412)
- ita (118)
- amharic (12)
- ancioloke (18)
- Arabic (18)
- Arabic, Moroccan Spoken (481)
- more...

ORGANISATION
- Académie marquisienne (Cubah Yq)
- CERTEC / Humanistlaboratoriet (1)
Easy as Geography
–
Google Earth Overlays
• geographic navigation seems like an obvious approach for novice users
• Google Earth is a popular, freely available tool
• KML format is widely used and easily convertible
• place marks for
  – linguistic archives
  – language sites
  – entry point for sets of resource bundles
• place marks can be enriched with introductory texts, photos and direct links to the MPI archive
Back to the Communities

Community Web Portals
• attractive graphical design
• tailor-made for a specific language community
• open source CMS as back end
• communication with the MPI archive through REST interface
Dane-zaa Community Portal

This page is created for the members of the Dane-zaa community to facilitate the use of the archive collected by the DoSeS team together with the elders.

### Stories
- Personal history
- Traditional stories
- Animals
- Place

### Learn about ...
- Drum
- Food and cooking
- Handgames
- Horses
- Moccassins
- Moosehide
- Preparing meat

### Materials
- Movies
- Clickables
- Dictionary
- Phrasebook
- Calendar
- Alphabet
- Posters

### The Archive
- Searching
- Navigating
- Download
- Tours
- Google Earth
- Studies
Accessing the Data

Word for word navigation
- LEXUS
• web-based tool for creating & editing lexica
• complies to LMF ISO standard
• but still very flexible
• lexical entries can be enriched with multimedia (video, audio, photos)
• multimedia can simply be linked from the archive
dada
(N)

fish sp (black spot sea perch)

more than

Pimtyó pumu y:a dada doo ya,
more than 10 people were with Pimtyó
should (tentative)
Browsing Concepts
- ViCoS
• supplement to LEXUS
• also web-based
• users can link concepts
• a number of “universal” relation types are provided
• users can define their own culture-specific relations
• idea is to bring indigenous people onboard
• diverse data in the MPI archive
• different user groups want access
• (open) IMDI standard and REST interfaces make different ways of accessing possible
• MPI tries to create alternative access methods customized for specific groups
Archiving instance,
Max Planck Institute for Psycholinguistics

Archive managers: 3
Archive developers: 2
System manager: 1
Archiving software development: 4
Enrichment software development: 4

Archive for language data:
40 Terabyte of data
400,000 archived objects