Technology in language documentation

Jacquelijn Ringersma

Max Planck Institute for Psycholinguistics

Language (archiving) technology

Language documentation:

Aim:

Maintain, consolidate or revitalize endangered languages

Through:

Creation of a representative, multipurpose and long-lasting record of languages

By:

Recording language events, speech and gesture in natural context etc.

Storing the resources in an organized, accessible and persistent archive

In this presentation:

Why archiving?

What is an organized, accessible and persistent archive?

Which technology is required (and offered by the MPI)

- Metadata tool
- Archive upload and access management tools
- Browsing, searching and accessing the resources
- Enrichments with ELAN, LEXUS and ViCoS

Why archiving?

Misconceptions about archiving

1. Your stuff is buried here and gone forever

(from Andrew Garret
Berkeley Archives)

Why archiving?

Misconceptions about archiving

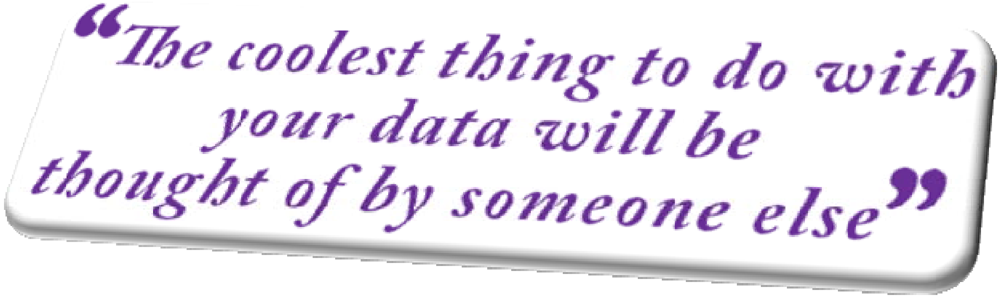
1. Your stuff is buried here and gone forever
2. Other linguists will take advantage of your hard work and take away your good ideas

(from Andrew Garret
Berkeley Archives)

Why archiving?

But the actual truth is that:

1. Other linguists do not really care about your work
2. The people who do care are the members of the speech communities – and they care about it in a different way than you do



"The coolest thing to do with your data will be thought of by someone else"

(attributed to Rufus Pollack)

Why archiving?

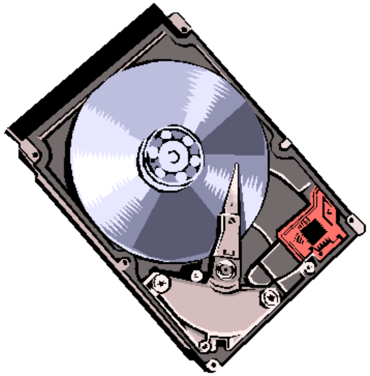
Is there a danger that we loose digital data?

YES,

UNESCO: 80% of our recordings is endangered

How much of your data and files on the notebook is organized, backed-up?

How long can media and formats be accessed?



Why archiving?

Is there a danger that we loose the data?

YES, a few messages

Archive data into a trusted archive (long term preservation and accessibility)
Create high quality metadata so that you can find the way back to the data
Use open standards

MPI-PL archive:
Trusted archive,

ORGANIZED information
Continuous extension
Collaboration and interaction
Commentary and relation drawing (enrichment)
Supporting centre for cross-corpus and language work

Why archiving?

Correct conceptions about archiving

1. It requires discipline
2. It creates a bit of techno noise

What’s an organized, accessible and persistent archive

Task of the ‘archiving instance’

Organization of corpora or data following clear principles

Creation of a coherent and consistent archive
Store data in an accessible and persistent form (long-term)

Give access to data to different users, but protect data against unauthorized access

Adhere to code of conduct and adhere to ethical and legal issues
Provide tools to researchers

Language archiving technology

Clear principles: Data organization and access infrastructure

IMDI Metadata editor
Browser and search

Coherent, consistent and persistent: Data management

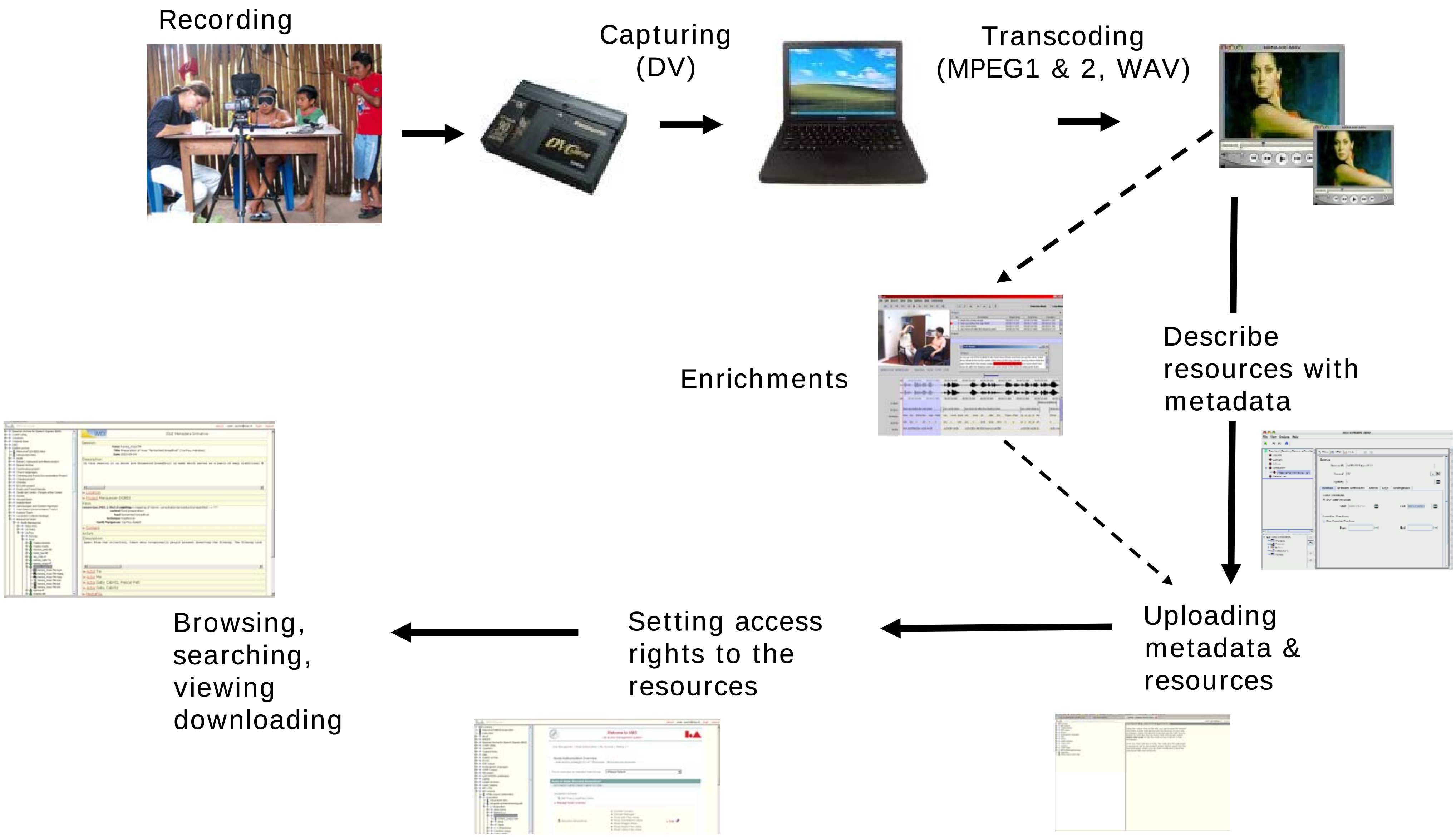
LAMUS Checks the content of the files, and file type check
Assigns a persistent identifier to the uploaded file
Allows the creation of corpus structures
Web based, easy to use

Safe access: Data access rights and protection

AMS All metadata in the archive is open
All resource access can be controlled by AMS (web based)
Users remain the owners and stay in control of the access
Setting of licenses and code of conducts

Tools required

Workflow and tool requirements:



Recording

Capturing (DV)

Transcoding (MPEG1 & 2, WAV)

Enrichments

Describe resources with metadata

Uploading metadata & resources

Setting access rights to the resources

Browsing, searching, viewing downloading

Tools required

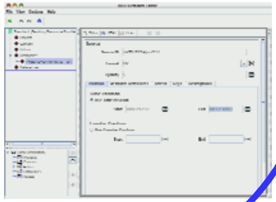
Workflow and tool requirements:

Recording

Capturing
(DV)

Transcoding
(MPEG1 & 2, WAV)

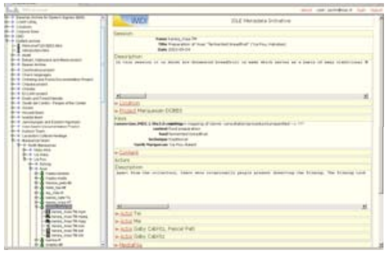
Describe
resources with
metadata



Uploading
metadata &
resources

Setting access
rights to the
resources

Browsing,
searching,
viewing
downloading



Tools required

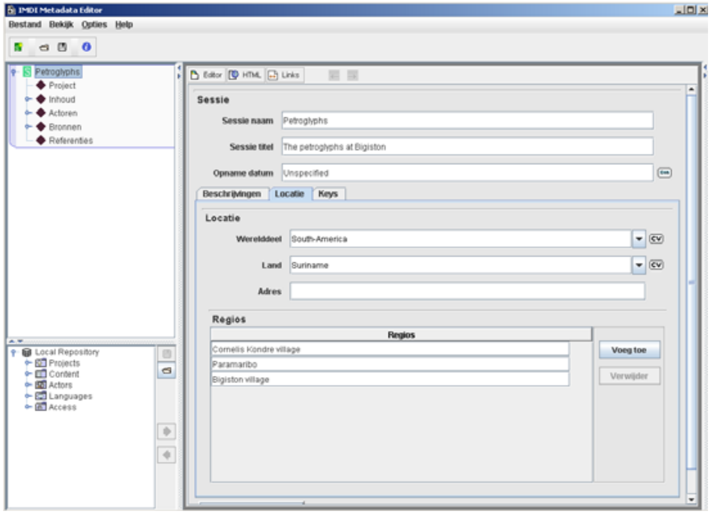
Describe resources with metadata

IMDI metadata

Metadata is data about data

Structured and machine readable

Elements describe the content of the resource files



Sessie

Sessie naam: Petroglyphs

Sessie titel: The petroglyphs at Bigiston

Opname datum: Unspecified

Locatie

Werelddeel: South-America

Land: Suriname

Adres:

Regios

Regios
Cornelis Kondre village
Paramaribo
Bigiston village

IMDI set:

General data: project, location

Content data: Genre, Interactivity, Modality, Language

Actor data: Age, gender, languages

Resource data: format, size

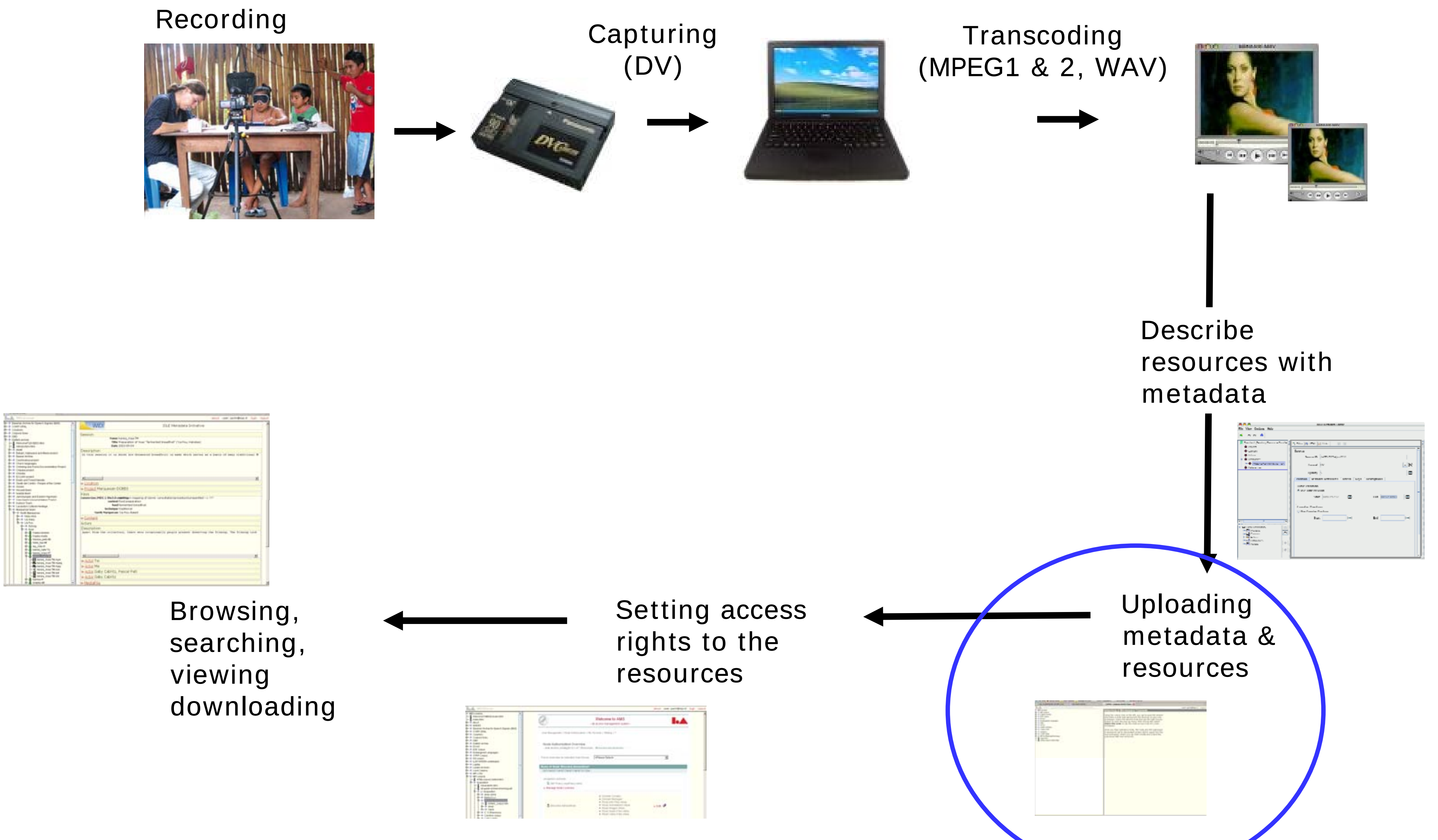
IMDI editor – downloadable from the LAT page

LAT – Language archiving technology

www.mpi-lat.eu

Tools required

Workflow and tool requirements:



Recording

Capturing (DV)

Transcoding (MPEG1 & 2, WAV)

Describe resources with metadata

Uploading metadata & resources

Setting access rights to the resources

Browsing, searching, viewing downloading

Tools required

Upload resources and metadata files with LAMUS

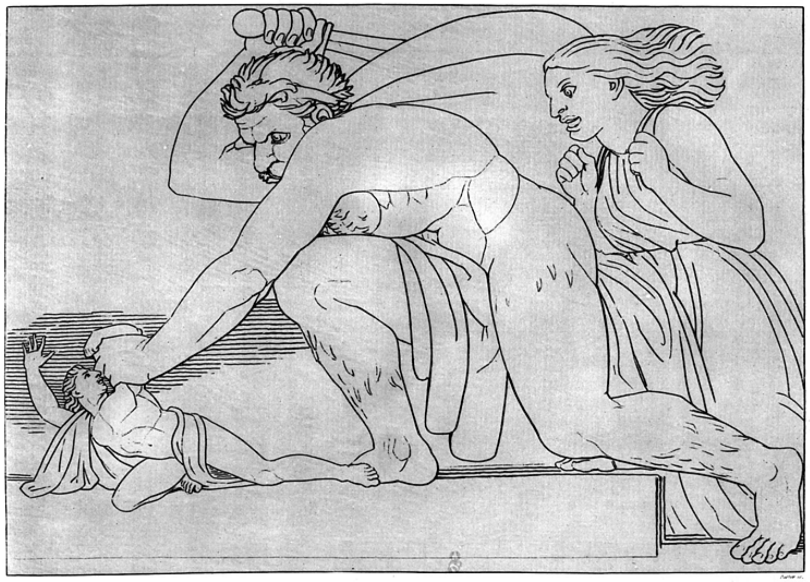
LAMUS

Checks the content of the files, and file type check

Assigns a persistent identifier to the uploaded file

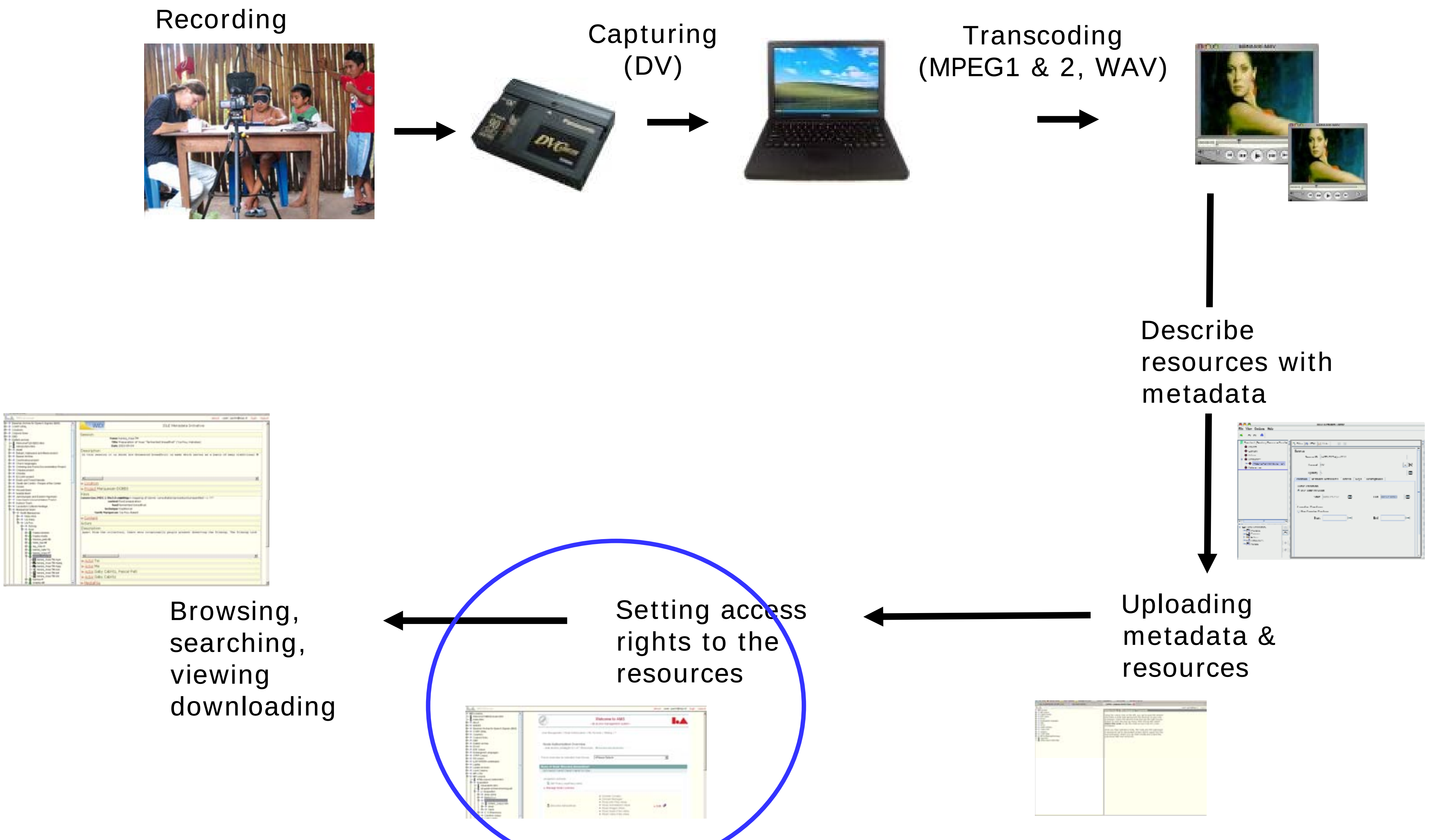
Allows the creation of corpus structures

Web based, easy to use



Tools required

Workflow and tool requirements:



Recording

Capturing (DV)

Transcoding (MPEG1 & 2, WAV)

Describe resources with metadata

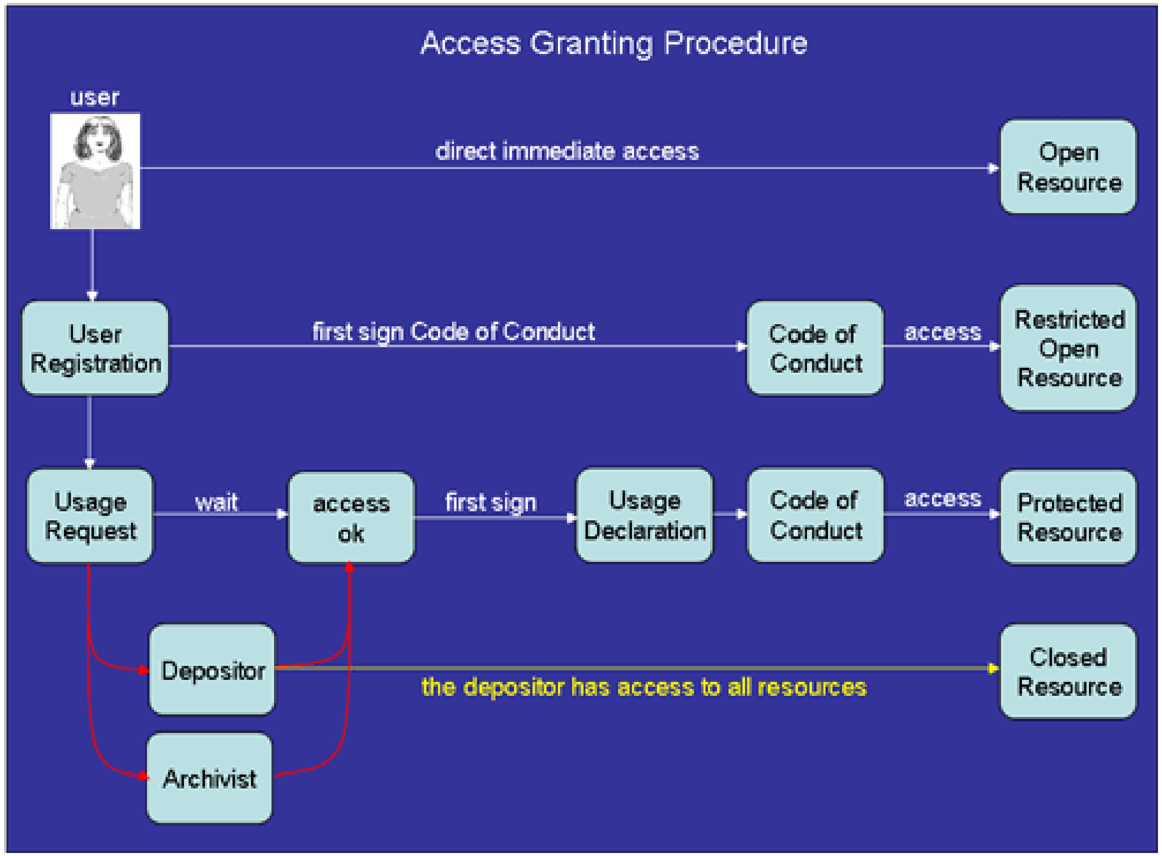
Uploading metadata & resources

Setting access rights to the resources

Browsing, searching, viewing downloading

Setting access rights: AMS

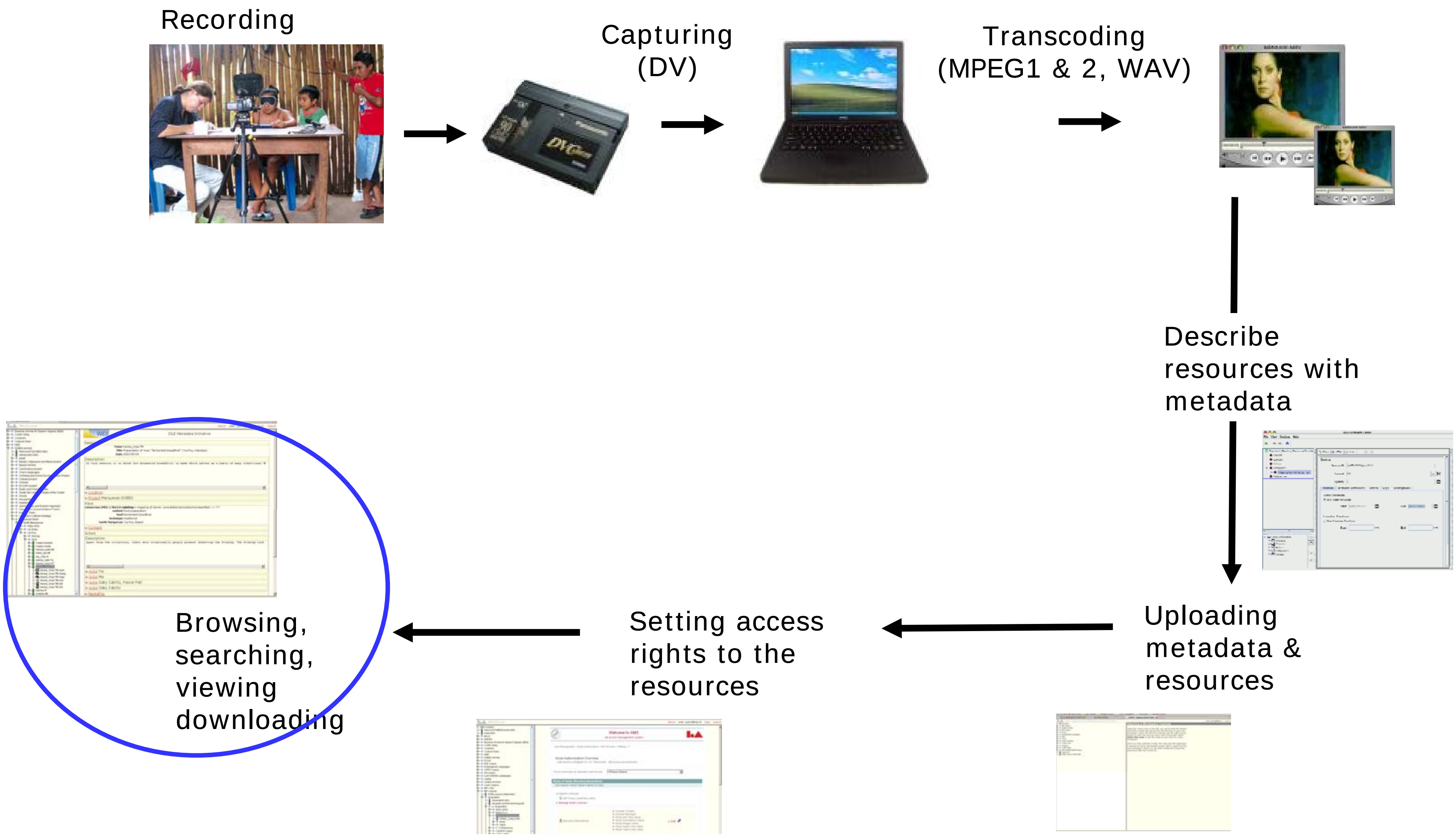
Access on resources: IPR, privacy, copyright etc.
(Metadata is always open!)



Default setting after upload

Tools required

Workflow and tool requirements:



Recording

Capturing (DV)

Transcoding (MPEG1 & 2, WAV)

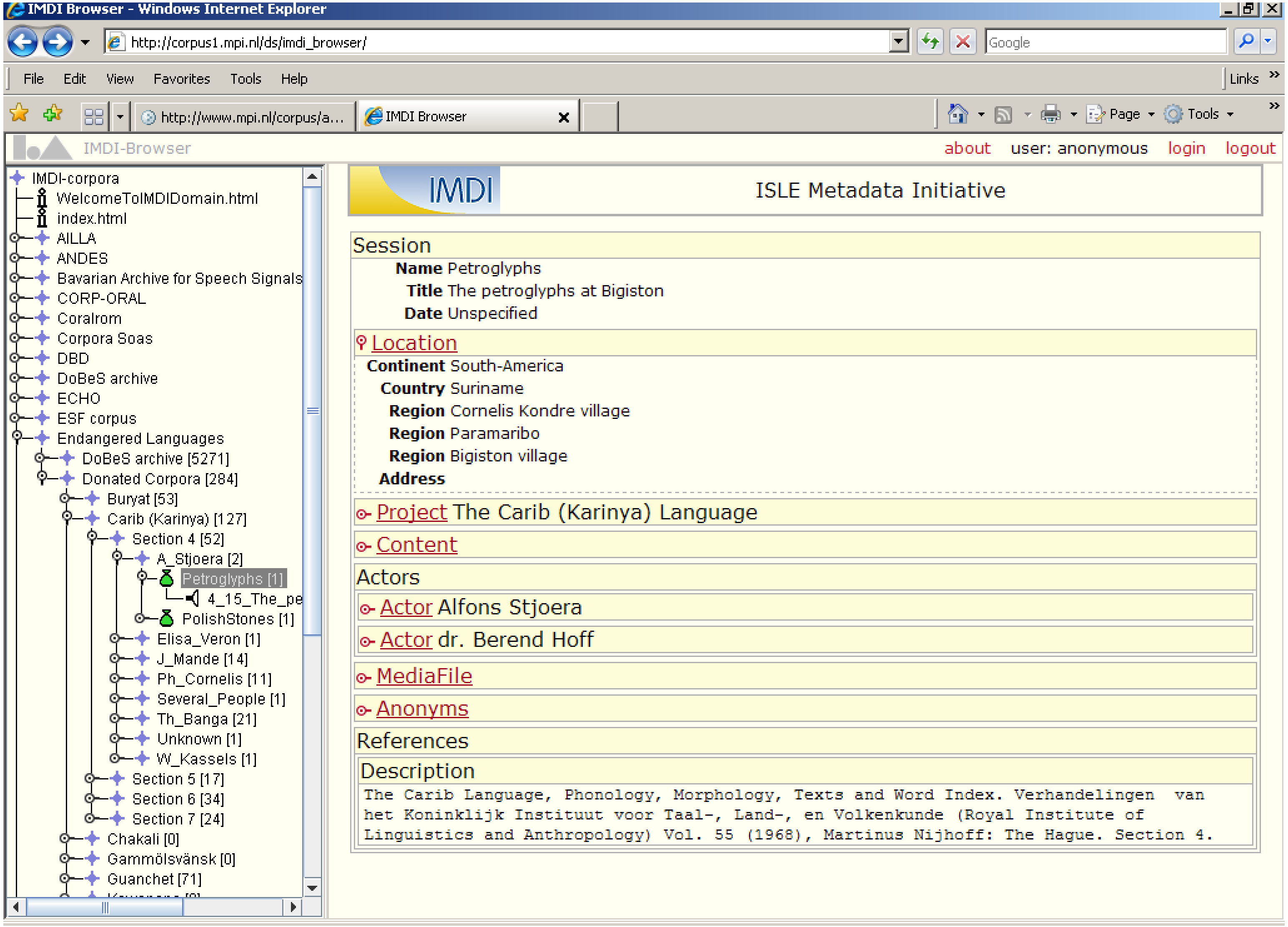
Describe resources with metadata

Uploading metadata & resources

Setting access rights to the resources

Browsing, searching, viewing downloading

Browsing the data: IMDI browser


IMDI Browser - Windows Internet Explorer
http://corpus1.mpi.nl/ds/imdi_browser/
Google
File Edit View Favorites Tools Help
Links
http://www.mpi.nl/corpus/a...
IMDI Browser
Page
Tools
IMDI-Browser
about
user: anonymous
login
logout
IMDI

- IMDI-corpora
  - WelcomeToIMDIDomain.html
  - index.html
  - AILLA
  - ANDES
  - Bavarian Archive for Speech Signals
  - CORP-ORAL
  - Coralrom
  - Corpora Soas
  - DBD
  - DoBeS archive
  - ECHO
  - ESF corpus
  - Endangered Languages
    - DoBeS archive [5271]
    - Donated Corpora [284]
      - Buryat [53]
      - Carib (Karinya) [127]
        - Section 4 [52]
          - A_Stjoera [2]
            - Petroglyphs [1]
              - 4_15_The_pe
            - PolishStones [1]
          - Elisa_Veron [1]
          - J_Mande [14]
          - Ph_Cornelis [11]
          - Several_People [1]
          - Th_Banga [21]
          - Unknown [1]
          - W_Kassels [1]
        - Section 5 [17]
        - Section 6 [34]
        - Section 7 [24]
      - Chakali [0]
      - Gammölsvänsk [0]
      - Guanchet [71]

ISLE Metadata Initiative

Session

Name Petroglyphs
Title The petroglyphs at Bigiston
Date Unspecified

Location

Continent South-America
Country Suriname
Region Cornelis Kondre village
Region Paramaribo
Region Bigiston village
Address

Project The Carib (Karinya) Language

Content

Actors

Actor Alfons Stjoera

Actor dr. Berend Hoff

MediaFile

Anonyms

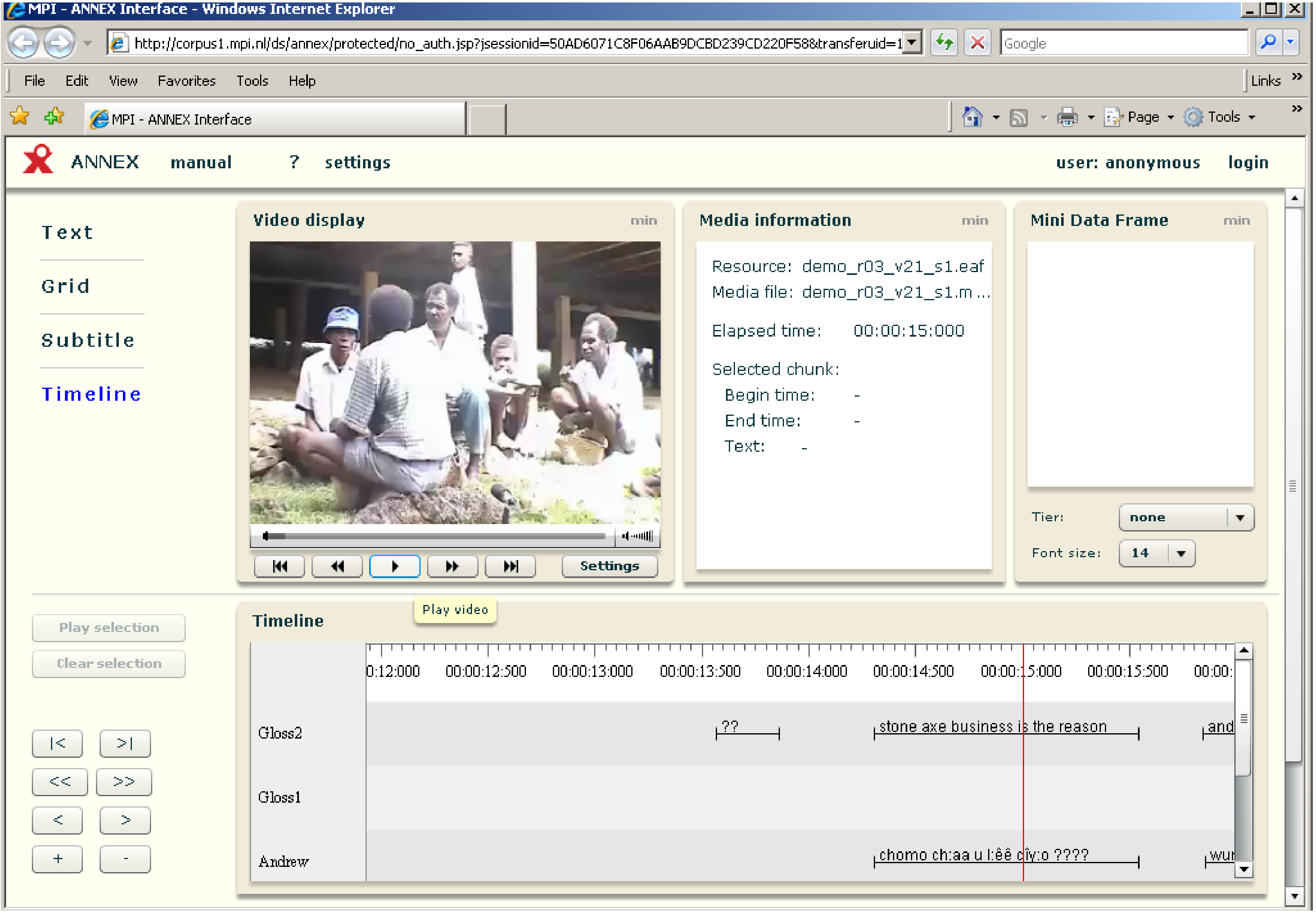
References

Description

The Carib Language, Phonology, Morphology, Texts and Word Index. Verhandelingen van het Koninklijk Instituut voor Taal-, Land-, en Volkenkunde (Royal Institute of Linguistics and Anthropology) Vol. 55 (1968), Martinus Nijhoff: The Hague. Section 4.

Tools required

Viewing the data: ANNEX viewer



http://corpus1.mpi.nl/ds/annex/protected/no_auth.jsp?jsessionid=50AD6071C8F06AAB9DCBD239CD220F58&transferuid=1

user: anonymous

Media information

Resource: demo_r03_v21_s1.eaf
Media file: demo_r03_v21_s1.m ...

Elapsed time: 00:00:15:000

Selected chunk:
Begin time: -
End time: -
Text: -

Tier: none

Font size: 14

Timeline

0:12:000 | 00:00:12:500 | 00:00:13:000 | 00:00:13:500 | 00:00:14:000 | 00:00:14:500 | 00:00:15:000 | 00:00:15:500 | 00:00:

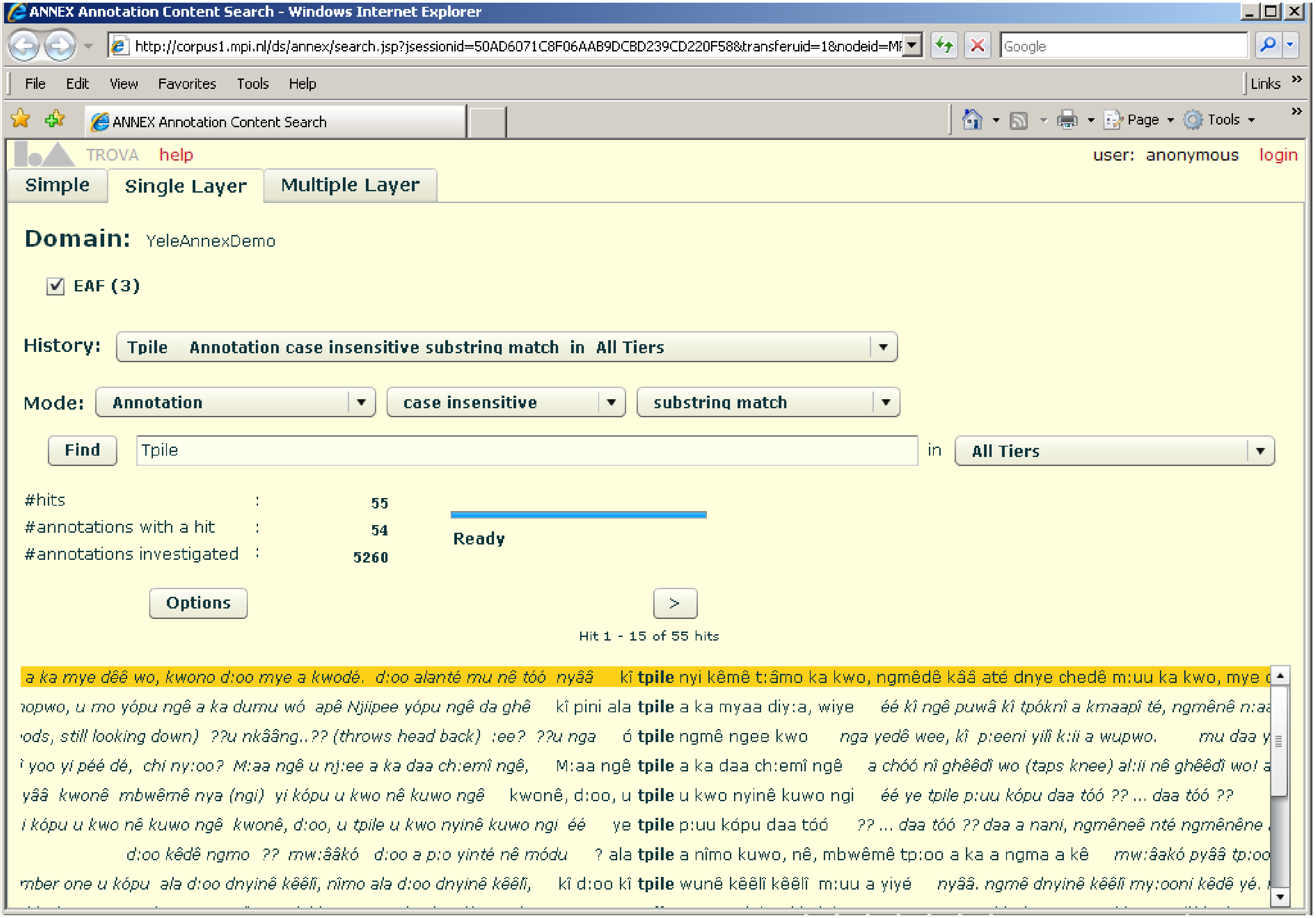
Gloss2: ?? | stone axe business is the reason | and

Gloss1

Andrew: chomo ch:aa u l:êê cíy:o ???? | wur

Tools required

Searching the metadata and content data: Search engines: IMDI and TROVA


ANNEX Annotation Content Search - Windows Internet Explorer
http://corpus1.mpi.nl/ds/annex/search.jsp?jsessionid=50AD6071C8F06AAB9DCBD239CD220F58&transferuid=1&nodeid=MP
Google
File Edit View Favorites Tools Help
Links
ANNEX Annotation Content Search
Page
Tools
TROVA help
user: anonymous login
Simple
Single Layer
Multiple Layer

Domain: YeleAnnexDemo

☑ EAF (3)

History: Tpile Annotation case insensitive substring match in All Tiers

Mode: Annotation | case insensitive | substring match

Find: Tpile in All Tiers

#hits : 55
#annotations with a hit : 54
#annotations investigated : 5260

Ready

Hit 1 - 15 of 55 hits

a ka mye dêê wo, kwono d:oo mye a kwodé. d:oo alanté mu nê tóó nyââ kî **tpile** nyi kêmê t:âmo ka kwo, ngmêdê kââ até dnye chedê m:uu ka kwo, mye c

ıopwo, u mo yópu ngê a ka dumu wó apê Njiipee yópu ngê da ghê kî pini ala **tpile** a ka myaa diy:a, wiye *éé kî ngê puwâ kî tpóknî a kmaapî té, ngmênê n:aa*

›ods, still looking down) ??u nkââng..?? (throws head back) :ee? ??u nga ó **tpile** ngmê ngee kwo *nga yedê wee, kî p:eeni yilî k:ii a wupwo. mu daa y*

ì yoo yi péé dé, chi ny:oo? M:aa ngê u nj:ee a ka daa ch:emî ngê, M:aa ngê **tpile** a ka daa ch:emî ngê *a chóó nî ghêêdî wo (taps knee) al:ii nê ghêêdî wo! a*

yââ kwonê mbwêmê nya (ngi) yi kópu u kwo nê kuwo ngê kwonê, d:oo, u **tpile** u kwo nyinê kuwo ngi *éé ye tpile p:uu kópu daa tóó ?? ... daa tóó ??*

i kópu u kwo nê kuwo ngê kwonê, d:oo, u tpile u kwo nyinê kuwo ngi éé ye **tpile** p:uu kópu daa tóó *?? ... daa tóó ?? daa a nani, ngmêneê nté ngmênêne .*

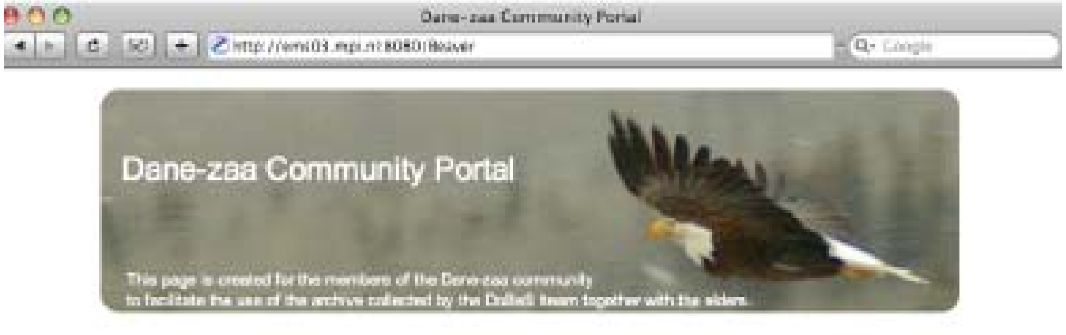
d:oo kêdê ngmo ?? mw:ââkó d:oo a p:o yinté nê módu ? ala **tpile** a nîmo kuwo, nê, mbwêmê tp:oo a ka a ngma a kê *mw:âakó pyââ tp:oo*

mber one u kópu ala d:oo dnyinê kêêlî, nîmo ala d:oo dnyinê kêêlî, kî d:oo kî **tpile** wunê kêêlî kêêlî m:uu a yiyé *nyââ. ngmê dnyinê kêêlî my:ooni kêdê yé.*

Tools required

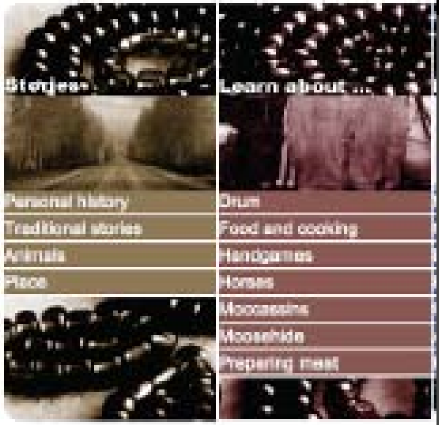
Different ways of accessing:

Community portals



Dane-zaa Community Portal

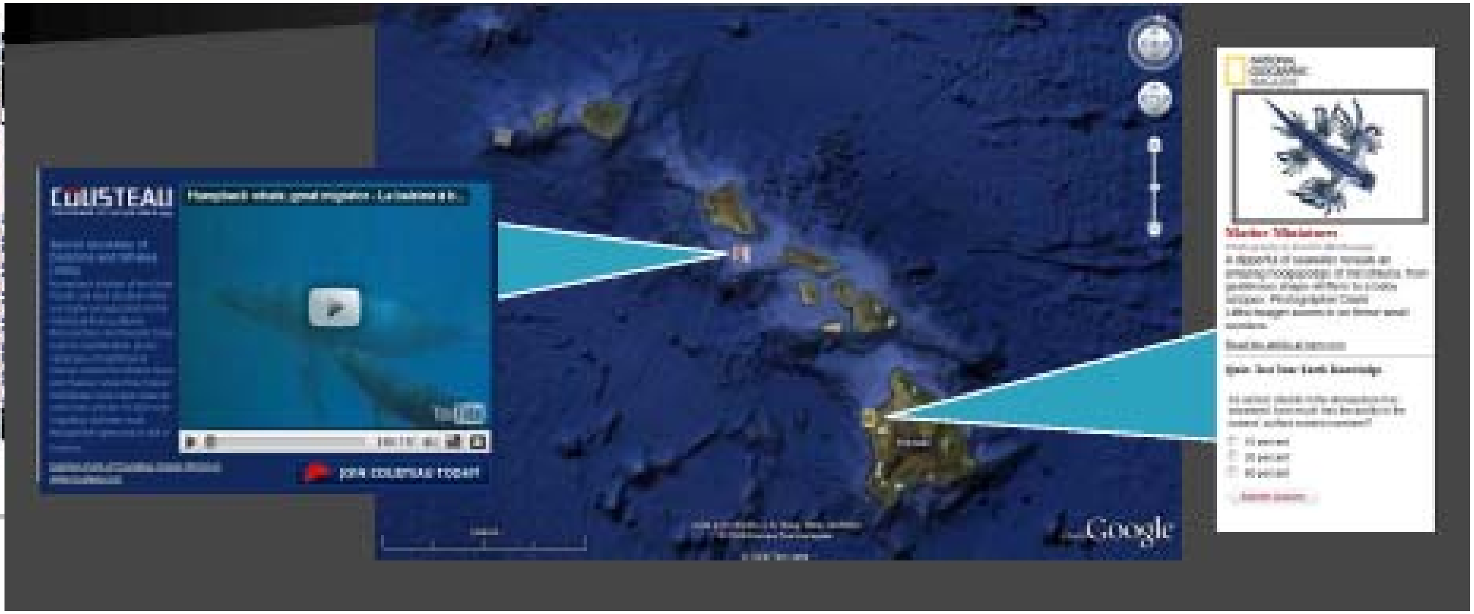
This page is created for the members of the Dane-zaa community to facilitate the use of the archive collected by the DoBeS team together with the elders.



- Personal history
- Traditional stories
- Animals
- Place

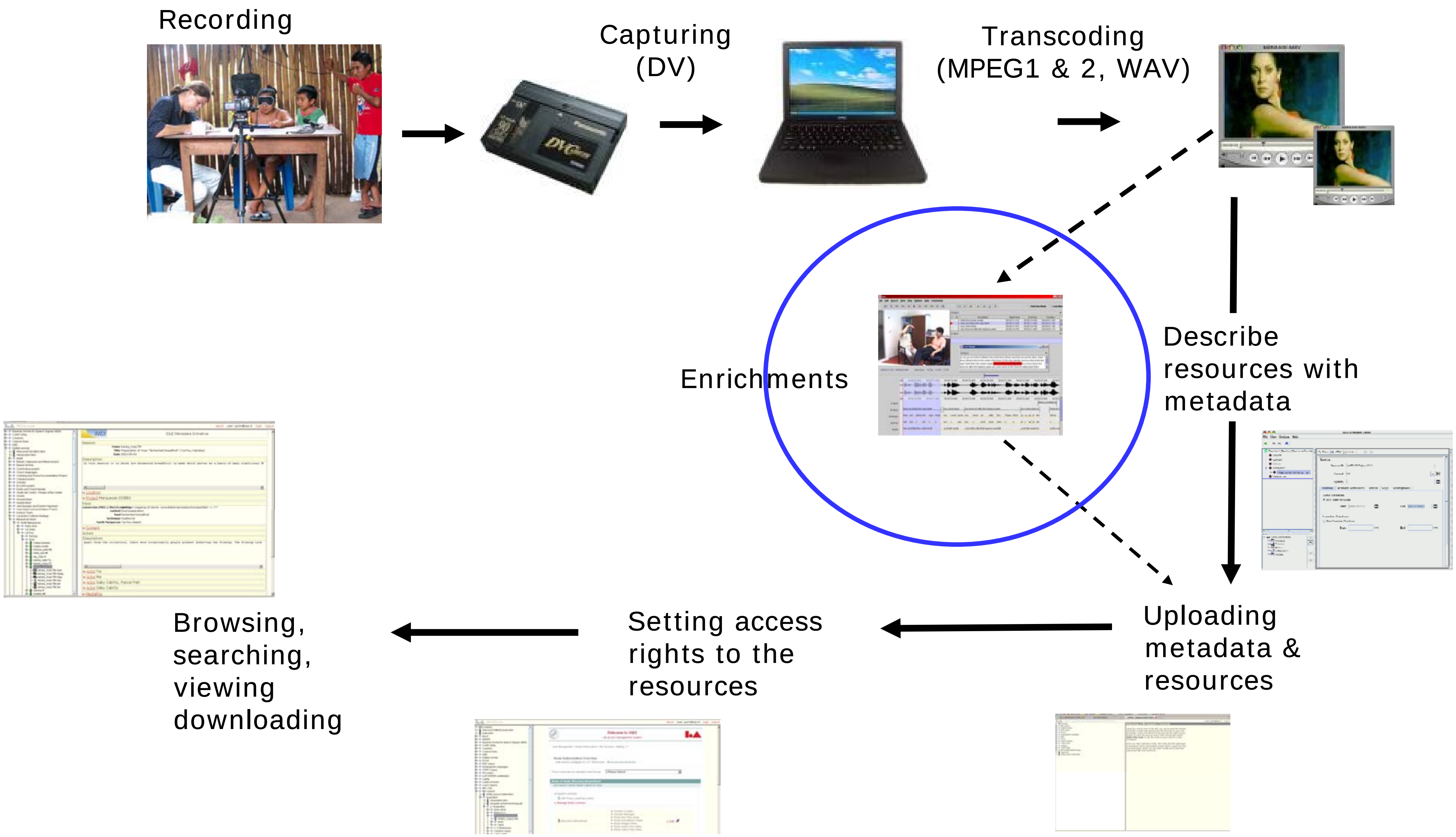
- Drum
- Food and cooking
- Handgames
- Horses
- Moccasins
- Moosehide
- Preparing meat

Google Earth Layers



Tools required

Workflow and tool requirements:



Recording

Capturing (DV)

Transcoding (MPEG1 & 2, WAV)

Enrichments

Describe resources with metadata

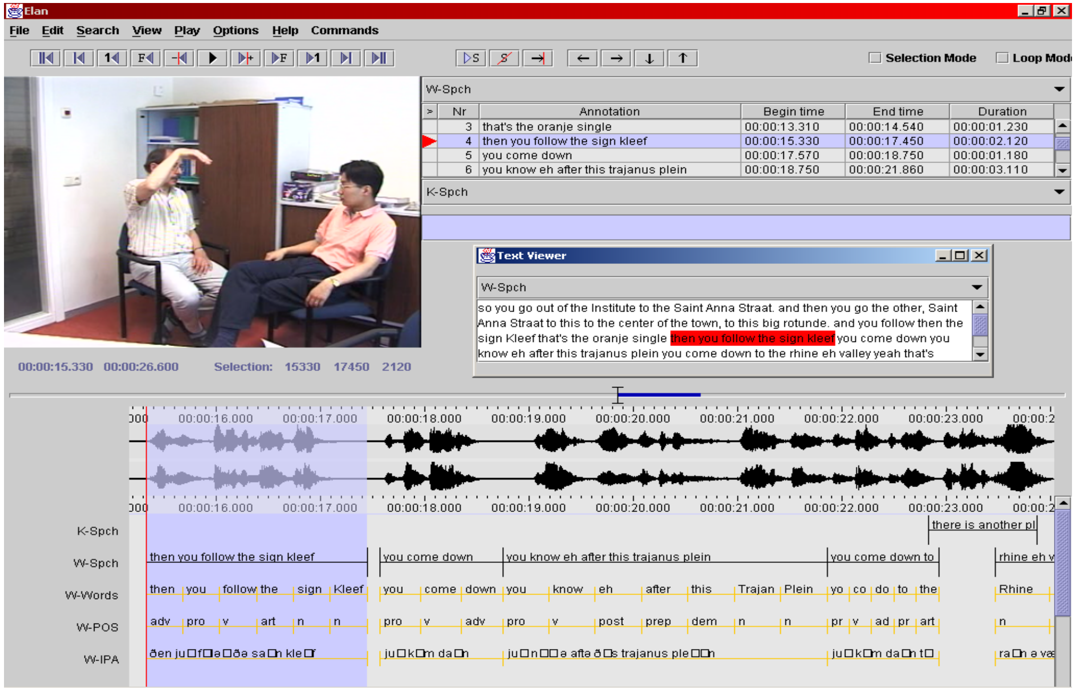
Uploading metadata & resources

Setting access rights to the resources

Browsing, searching, viewing downloading

Enrichments

- ELAN: annotating video and audio resources



W-Spch

>	Nr	Annotation	Begin time	End time	Duration
	3	that's the oranje single	00:00:13.310	00:00:14.540	00:00:01.230
▶	4	then you follow the sign kleef	00:00:15.330	00:00:17.450	00:00:02.120
	5	you come down	00:00:17.570	00:00:18.750	00:00:01.180
	6	you know eh after this trajanus plein	00:00:18.750	00:00:21.860	00:00:03.110

K-Spch

Text Viewer

W-Spch

so you go out of the Institute to the Saint Anna Straat. and then you go the other, Saint Anna Straat to this to the center of the town, to this big rotunde. and you follow then the sign Kleef that's the oranje single then you follow the sign kleef you come down you know eh after this trajanus plein you come down to the rhine eh valley yeah that's

K-Spch: there is another pl

W-Spch: then you follow the sign kleef | you come down | you know eh after this trajanus plein | you come down to | rhine eh v

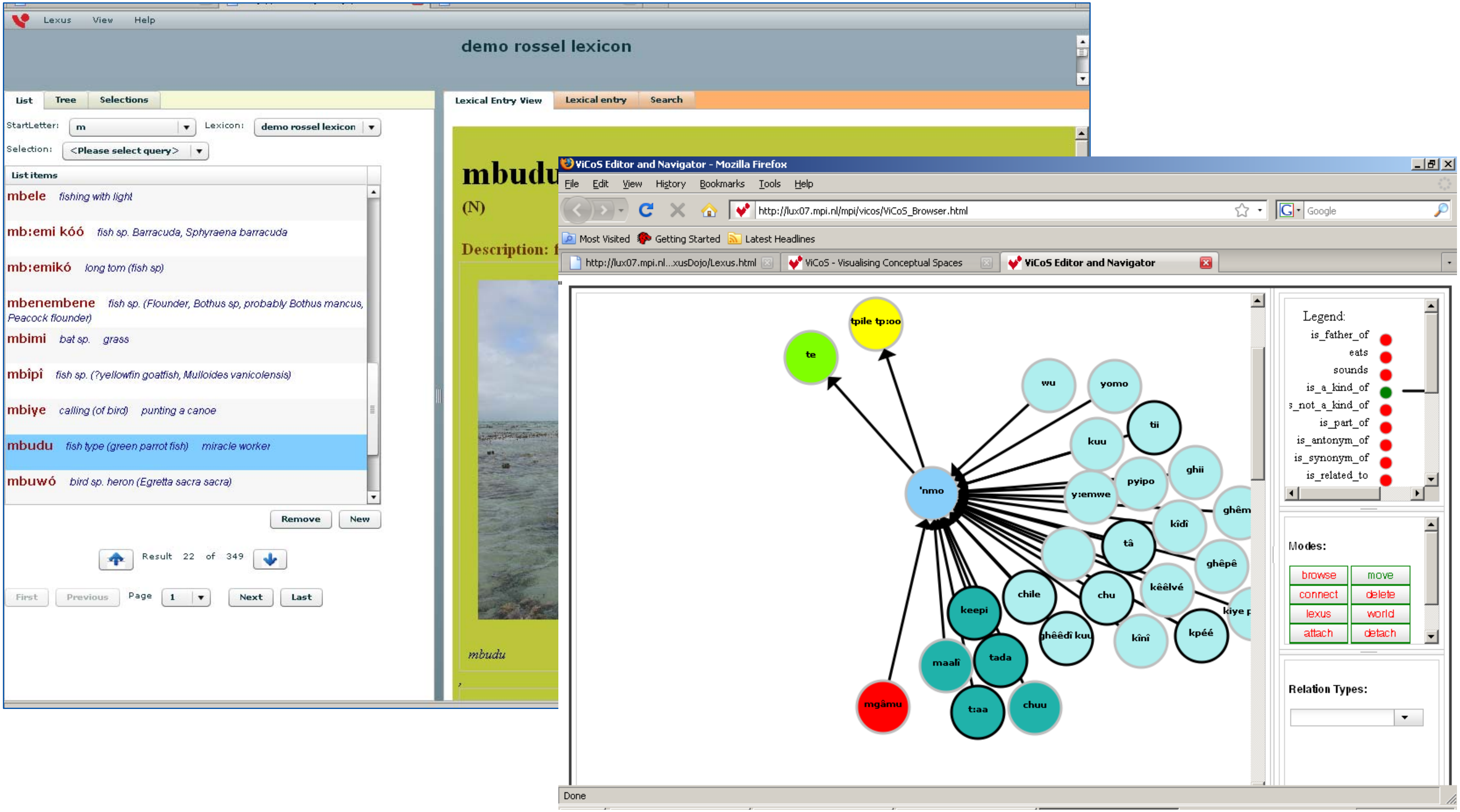
W-Words: then | you | follow the | sign | Kleef | you | come | down | you | know | eh | after | this | Trajan | Plein | yo | co | do | to | the | Rhine

W-POS: adv | pro | v | art | n | n | pro | v | adv | pro | v | post | prep | dem | n | n | pr | v | ad | pr | art | n

W-IPA: ðen ju□f□lə□ðə sa□n kle□f | ju□k□m da□n | ju□n□□ə aftə ð□s trajanus ple□□n | ju□k□m da□n t□ | ra□n ə væ

Enrichments

LEXUS & ViCoS: Web based lexicon tool, multimedia encyclopedia



Lexus View Help

demo rossel lexicon

List | Tree | Selections

StartLetter: m Lexicon: demo rossel lexicon

Selection: <Please select query>

List items

mbele	fishing with light
mb:emi kóó	fish sp. Barracuda, Sphyraena barracuda
mb:emikó	long tom (fish sp)
mbenembene	fish sp. (Flounder, Bothus sp, probably Bothus mancus, Peacock flounder)
mbimi	bat sp. grass
mbîpî	fish sp. (?yellowfin goatfish, Mulloides vanicolensis)
mbiye	calling (of bird) punting a canoe
mbudu	fish type (green parrot fish) miracle worker
mbuwó	bird sp. heron (Egretta sacra sacra)

Remove New

Result 22 of 349

First Previous Page 1 Next Last

Lexical Entry View | Lexical entry | Search

mbudu

(N)

Description:

mbudu

ViCoS Editor and Navigator - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://lux07.mpi.nl/mpi/vicos/ViCoS_Browser.html

Most Visited Getting Started Latest Headlines

http://lux07.mpi.nl...xusDojo/Lexus.html | ViCoS - Visualising Conceptual Spaces | ViCoS Editor and Navigator

Legend:

is_father_of	●
eats	●
sounds	●
is_a_kind_of	●
s_not_a_kind_of	●
is_part_of	●
is_antonym_of	●
is_synonym_of	●
is_related_to	●

Modes:

browse	move
connect	delete
lexus	world
attach	detach

Relation Types:

Done

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

Technology training

Training sessions:

We regularly organize training sessions on:

Audio and video handling

Archiving technology

Enrichment of data

We do welcome participants from other than DoBeS or MPI projects

Its 4-5 days in one week

Interested: please check our MPI website/Events section (www.mpi.nl/events/)

Contact:

jacquelijn.ringersma@mpi.nl

paul.trilsbeek@mpi.nl