

# DELAMAN / DAM-LR - the vision -

Digital Endangered Languages and Music Archives Network

Distributed Access Management for Language Resources  
(EU – Project started at 1.1.05)

Peter Wittenburg  
MPI for Psycholinguistics

- just 5 years ago that we started in our discipline speaking about
  - large digital online collections
  - standardizing the formats
  - open metadata to come to browsable and searchable domains
  - using open metadata to create well-organized archives
- LREC Athens 2000
  - first workshop on these issues
  - start of the ISLE project (linguistic concepts, lexicon, metadata, ...)
  - start of the work on the IMDI metadata infrastructure
- in late 2000 also first LDC workshop with OLAC as focus
- this is very short time when you want to convince a community

- have “large” on-line digital archives/collections/Digital Libraries
  - MPI ~40.000 session bundles (> 100.000 objects) / ~11 TB
  - DOBES housed at MPI ~1.500 session bundles/ 1500 h
  - AILLA archive
  - PARADISEC archive
  - Lund corpus archive
  - also in HLT domain larger data centers
  - also “traditional” archives (Phonogramm Archiv, NAA, ...)
  - etc
- idea of web visibility and online accessibility spreads
- necessity of central data collection and preservation spreads



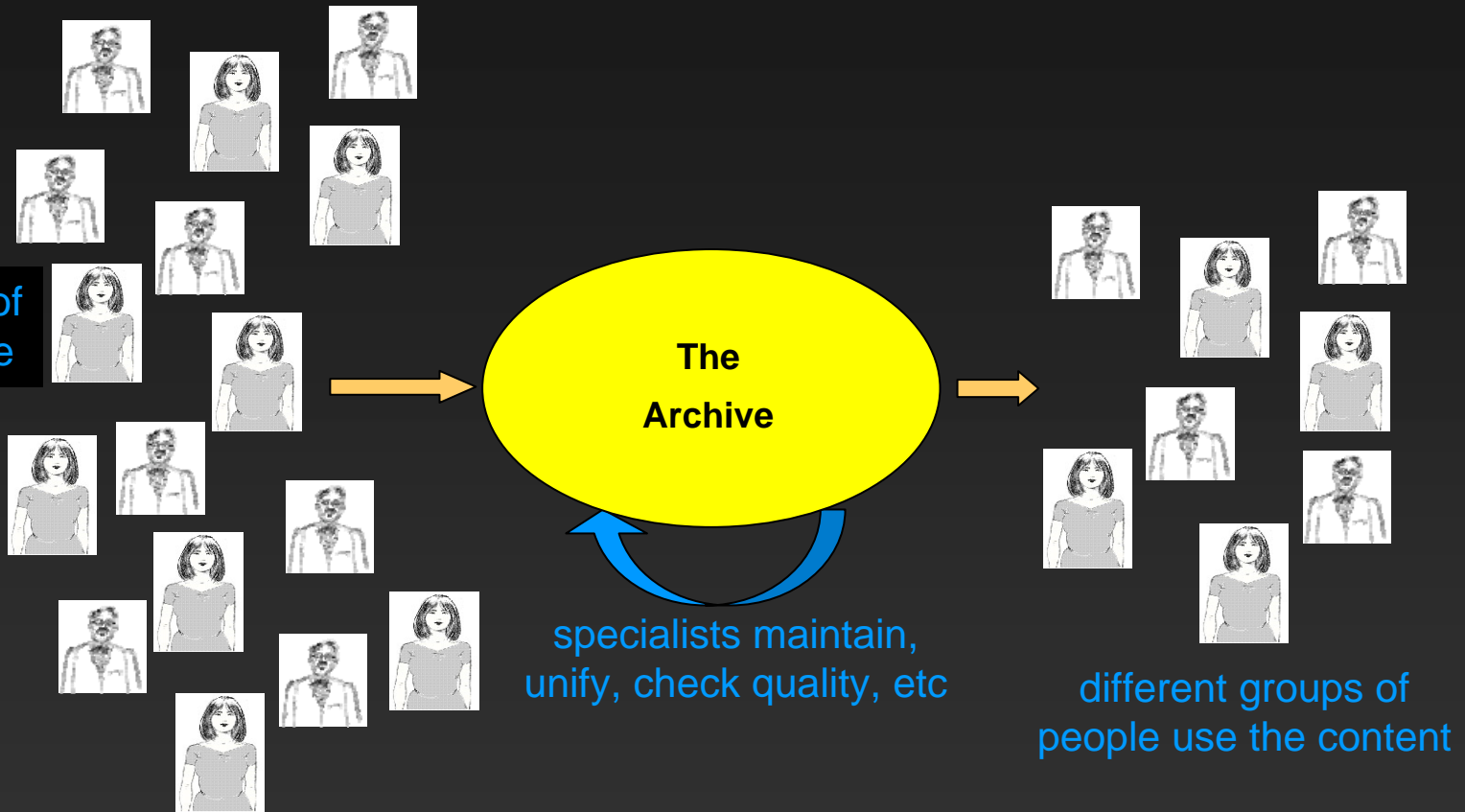
- much evangelization and agreement about standards
- “everyone” agrees with XML, UNICODE and linear PCM
- “everyone” understands the relevance of schemas to make linguistic structure and encoding explicit
- wrt JPEG and MPEG we are shooting on a moving target, but don’t yet have real alternatives

- interoperability is still a dream however ...
  - have metadata gateways in our discipline (OLAC-IMDI)
  - increasingly often tools are producing correct XML, UNICODE, ...
  - have filters for character encodings and formats although we miss well-designed and comprehensive services
  - have started with ontology work to tackle the linguistic aspects
    - GOLD ontology from E-Meld
    - ISO TC37/SC4 Data Category Registry
    - TDS (Dutch Typology Project) meta-language
    - EAGLES/ISLE/TEI specifications
- we are at the beginning
- cannot speak yet about fully operational infrastructures but there are island tools like FIELD, LEXUS, ONTO-ELAN, ...

REACBSWDVCPSENW  
 MDOWNVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNISOFEENDANG  
 KNDFNGFLANGUAGES  
 REACBSWDVCPSENW  
 MDOWNVTQQAQSPOR

different groups of people contribute

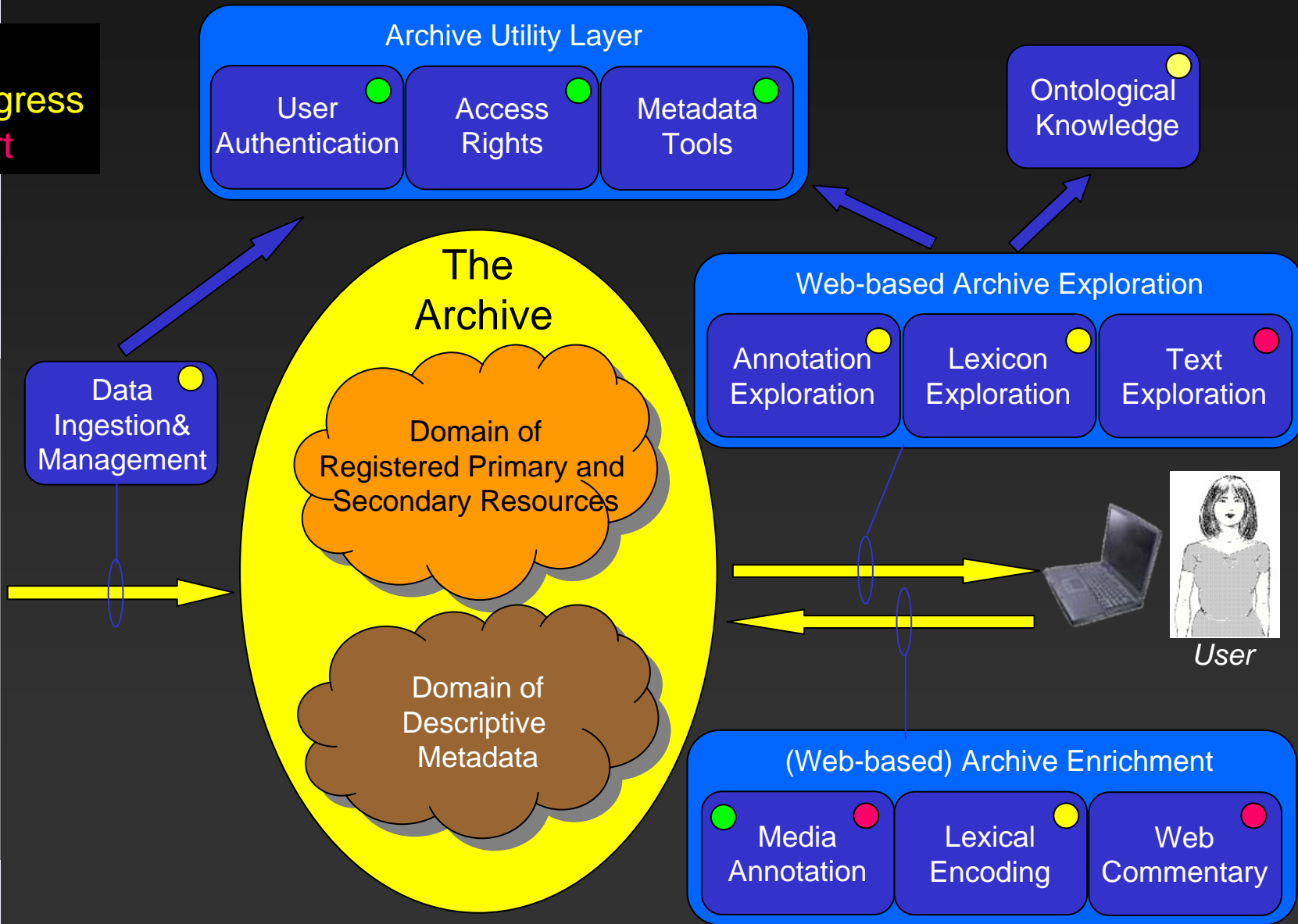
REACBSWDVCPSENW  
 MDOWNVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNISOFEENDANG  
 KNDFNGFLANGUAGES  
 REACBSWDVCPSENW  
 MDOWNVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNISOFEENDANG  
 KNDFNGFLANGUAGES  
 REACBSWDVCPSENW  
 MDOWNVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNISOFEENDANG  
 KNDFNGFLANGUAGES  
 REACBSWDVCPSENW  
 MDOWNVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNISOFEENDANG  
 KNDFNGFLANGUAGES



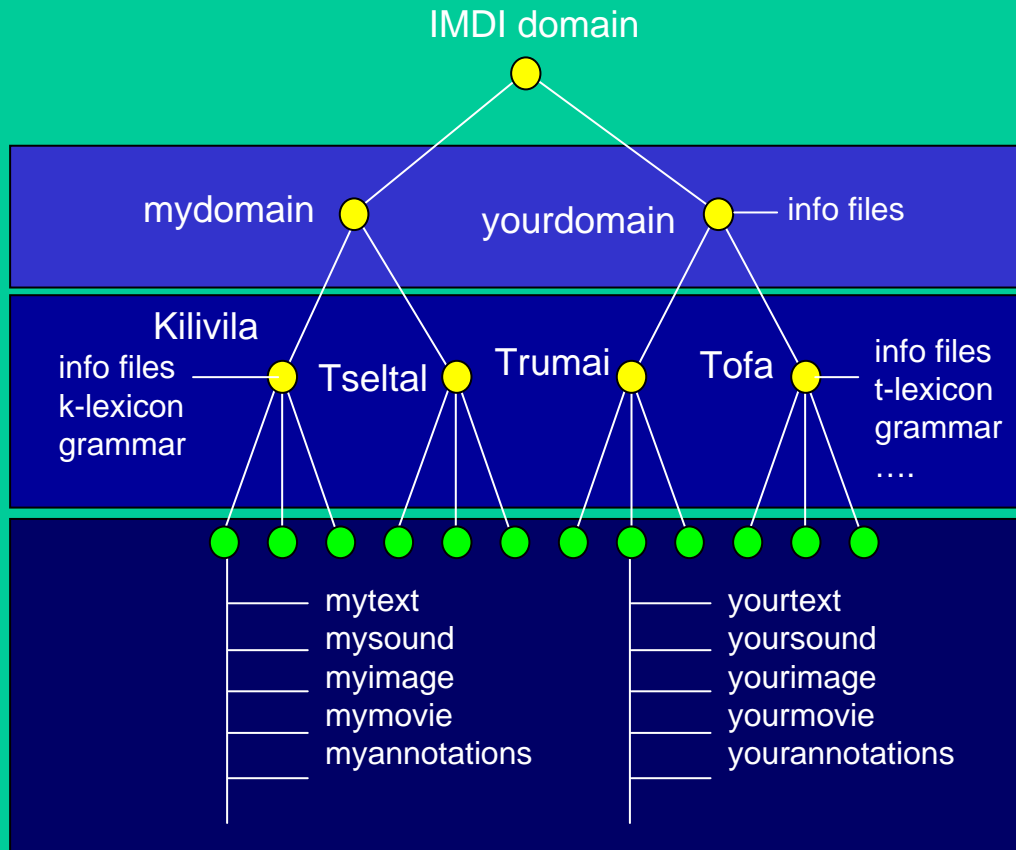
- at the MPI it is understood that the archive is the capital to build on
- in the DOBES programme the point to make results explicit and accessible
- only works if we don't have an "inert, dusty" archives
- language archives are dynamic!

# DOBES / MPI Archives as Example

- done
- in progress
- to start



## Archive Contents



- IMDI schema
- EAF schema
- myanno structure
- youranno structure
- LEXUS schema
- t-lex structure
- k-lex structure
- .....

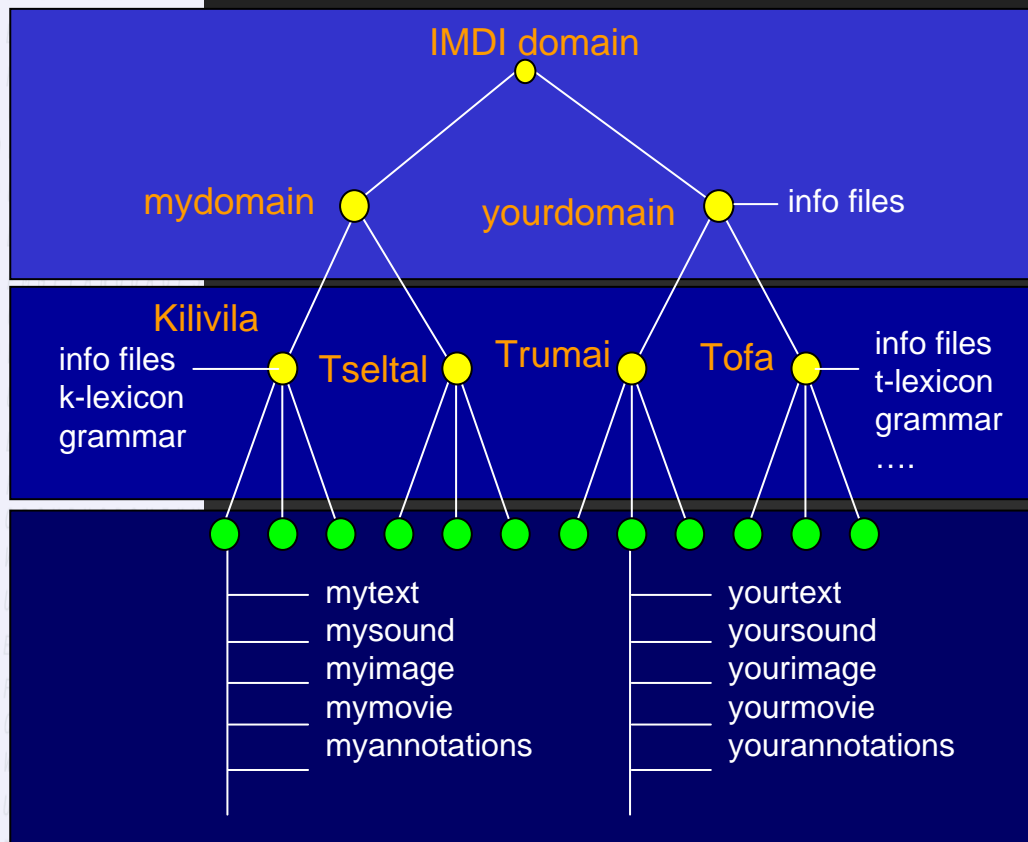
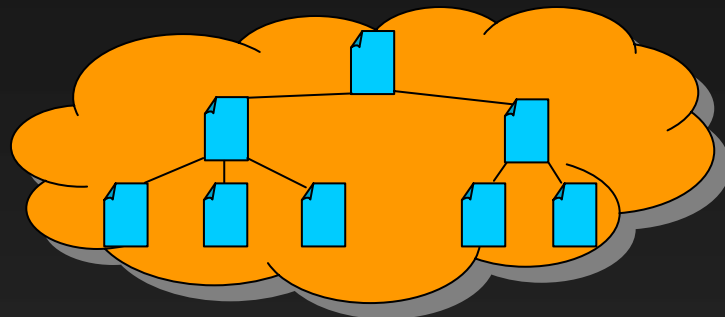


User



**Primary Resources:**  
 Texts  
 Images  
 Sound  
 Movies

- researcher free to define structure
- MD descriptions have to be correct (IMDI schema and CV)

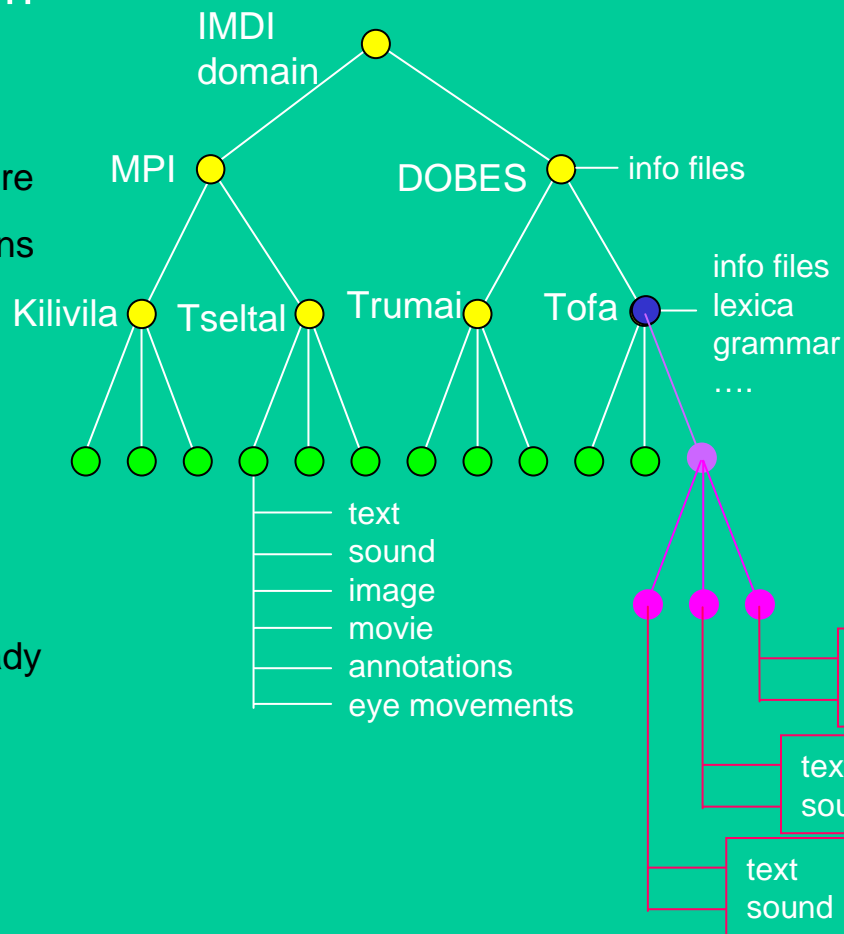


- fully distributed domain
- sufficient to register the root URL
- searching requires harvesting
- HTML browsing requires harvesting

## Resource Ingestion

1. upload/define structure
2. upload/define sessions
3. upload resources
4. link resources
5. define access policy
6. system to carry out checks

LAMUS Light almost ready



User

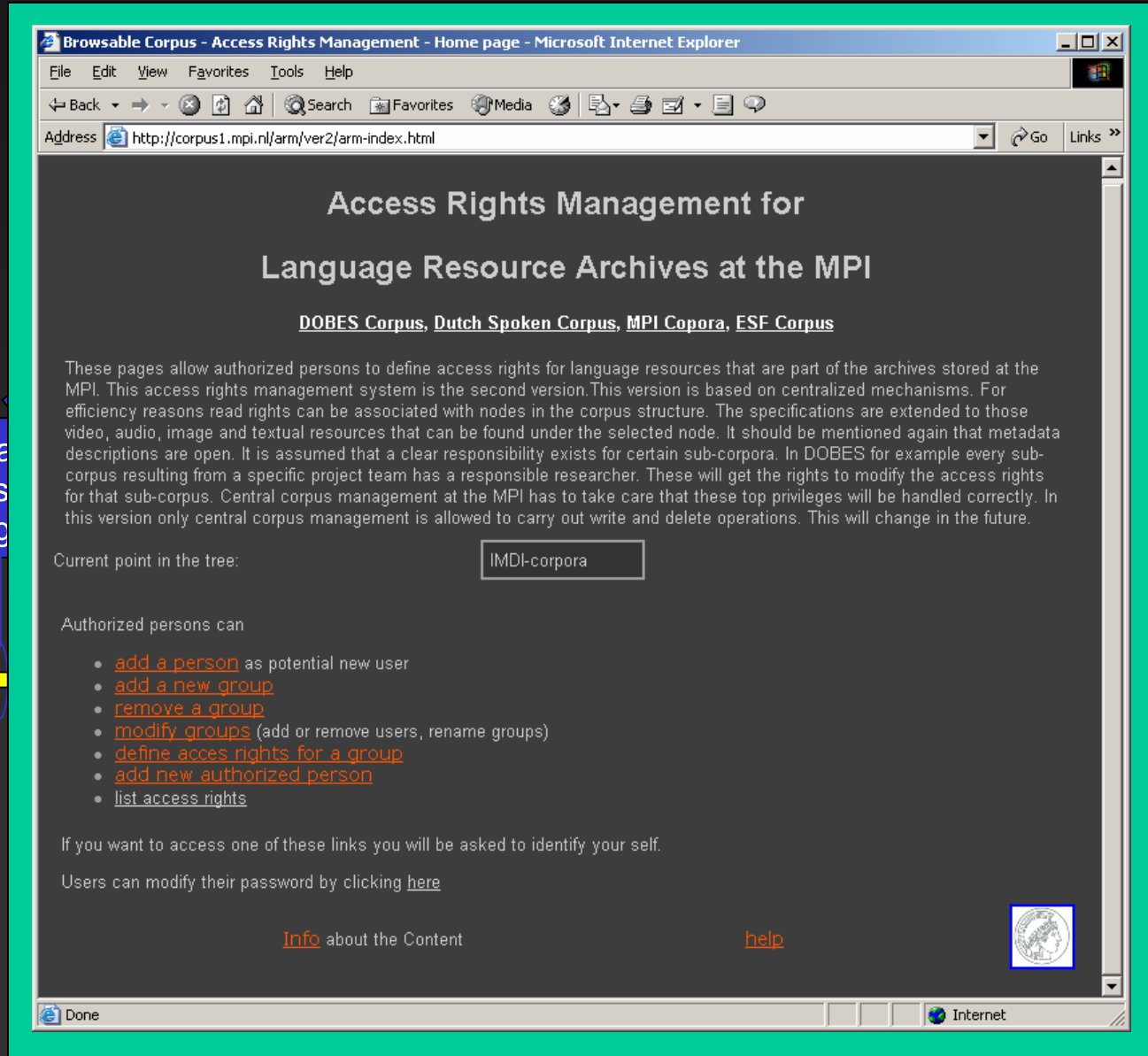


**Primary Resources:**  
 Texts  
 Images  
 Sound  
 Movies



REACBSWDVCPSEMEN  
 NDOWHVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDENGLANGUAGES  
 REACBSWDVCPSEMEN  
 NDOWHVTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDENGLANGUAGES  
 REACBSWDVCPSEMEN

Da  
 Inges  
 Manag



**Access Rights Management for Language Resource Archives at the MPI**

**DOBES Corpus, Dutch Spoken Corpus, MPI Copora, ESF Corpus**

These pages allow authorized persons to define access rights for language resources that are part of the archives stored at the MPI. This access rights management system is the second version. This version is based on centralized mechanisms. For efficiency reasons read rights can be associated with nodes in the corpus structure. The specifications are extended to those video, audio, image and textual resources that can be found under the selected node. It should be mentioned again that metadata descriptions are open. It is assumed that a clear responsibility exists for certain sub-corpora. In DOBES for example every sub-corpora resulting from a specific project team has a responsible researcher. These will get the rights to modify the access rights for that sub-corpora. Central corpus management at the MPI has to take care that these top privileges will be handled correctly. In this version only central corpus management is allowed to carry out write and delete operations. This will change in the future.

Current point in the tree: IMDI-corpora

Authorized persons can

- [add a person](#) as potential new user
- [add a new group](#)
- [remove a group](#)
- [modify groups](#) (add or remove users, rename groups)
- [define access rights for a group](#)
- [add new authorized person](#)
- [list access rights](#)

If you want to access one of these links you will be asked to identify your self.

Users can modify their password by clicking [here](#)

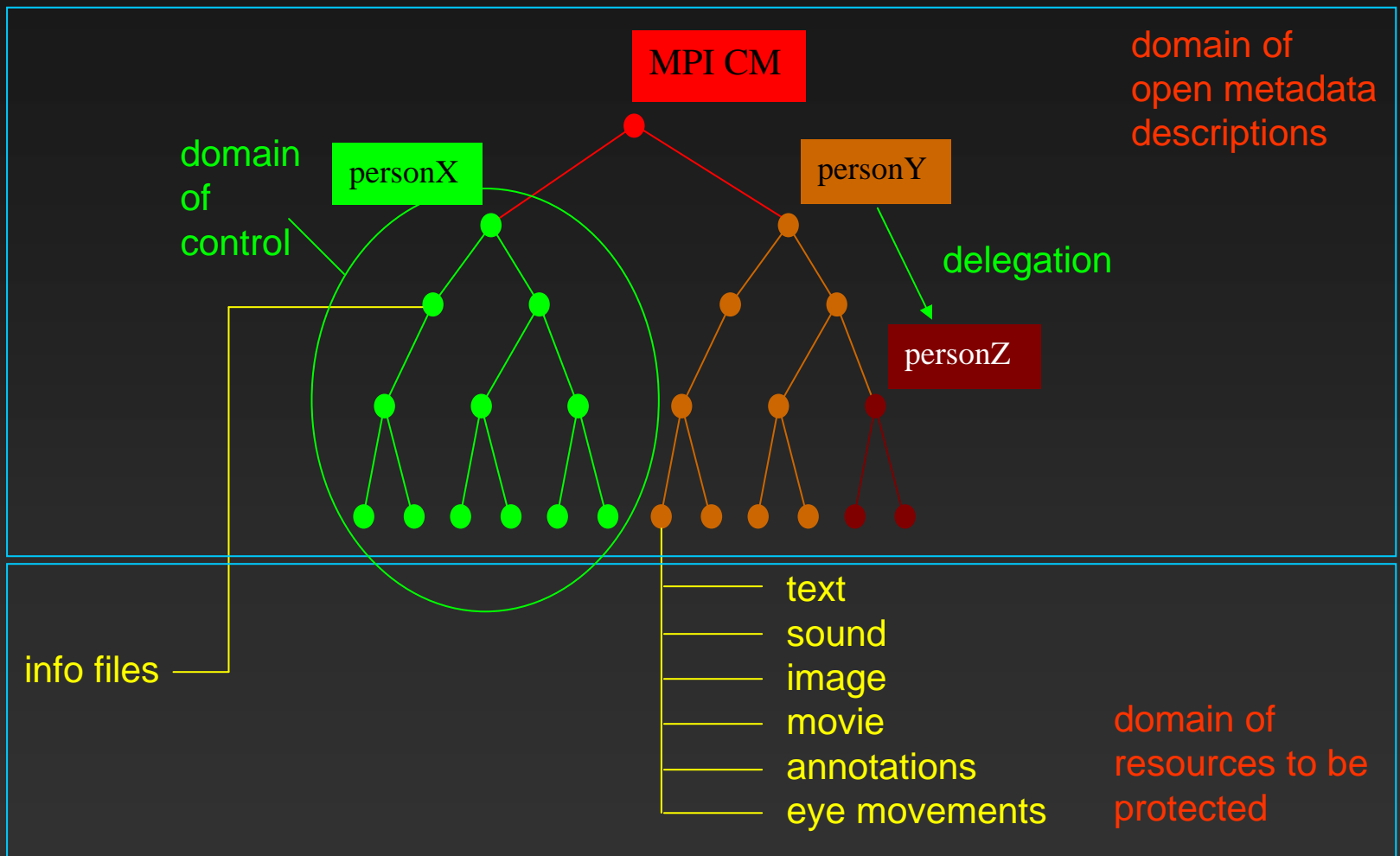
[Info](#) about the Content [help](#)



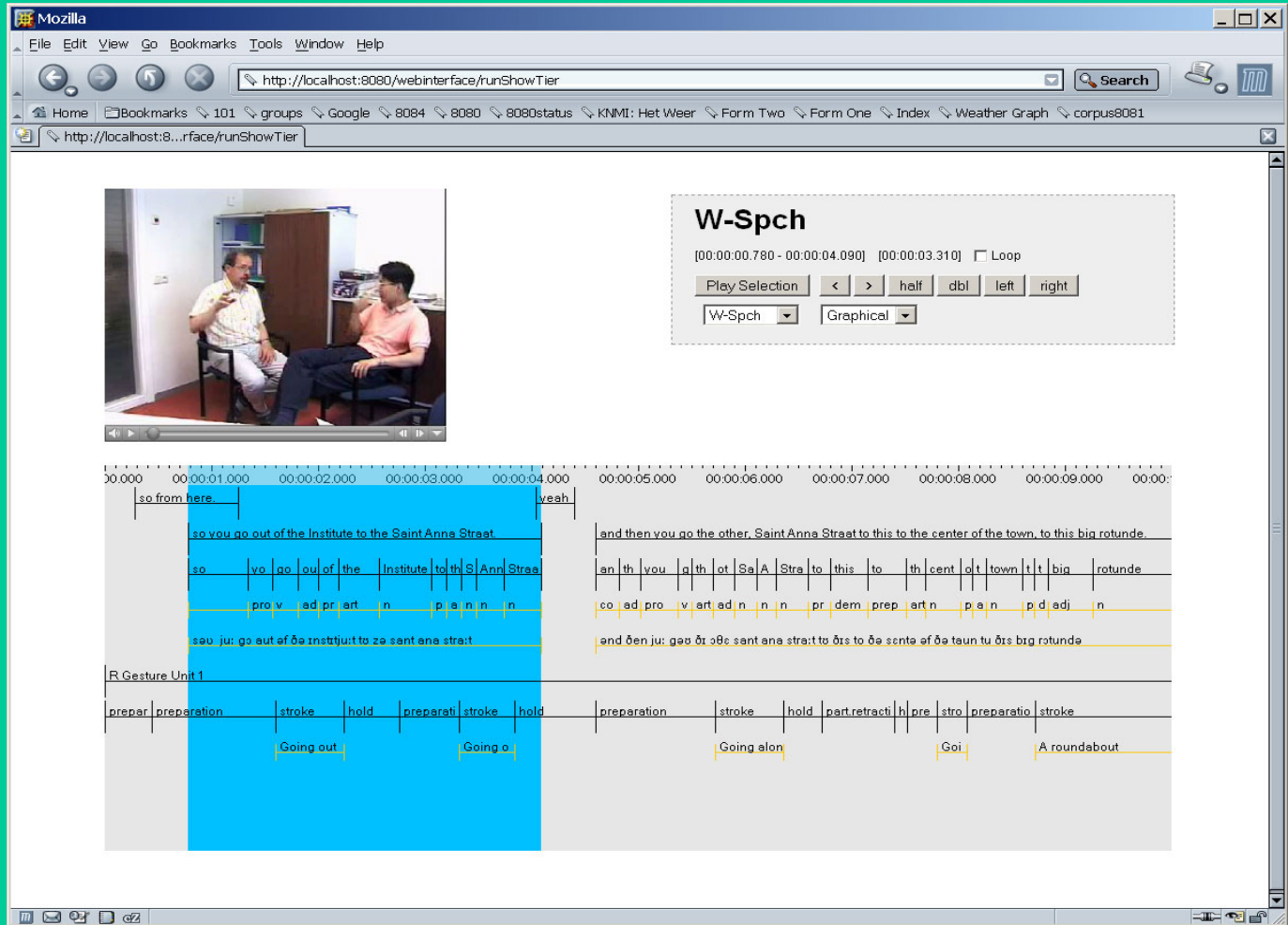
**Primary Resources:**  
 Texts  
 Images  
 Sound  
 Movies



User



- current solution is centralized – one database
- has delegation mechanism to make administration tractable
- association of declarations etc is possible
- powerful commands from any node to give rights to groups



The screenshot shows a Mozilla browser window displaying a video player. The video shows two men sitting and talking. The browser address bar shows `http://localhost:8080/webinterface/runShowTier`. The video player has a control bar with buttons for Play Selection, navigation, and volume. Below the video is a timeline with annotations. The annotations include English text, phonetic transcriptions, and gesture labels like "Going out" and "Going o".

**W-Spch**  
 [00:00:00.780 - 00:00:04.090] [00:00:03.310]  Loop  
 Play Selection < > half dbl left right  
 W-Spch Graphical

Timeline annotations (00:00:00.000 to 00:00:09.000):

- so from here | yeah
- 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.
- so | yo | go | ou | of | the | Institute | to | th | S | Ann | Straa | an | th | you | g | th | ot | Sa | A | Stra | to | this | to | th | cent | ot | town | t | big | rotunde
- pro v | ad pr | art | n | p | a | n | n | co | ad | pro | v | art | ad | n | n | n | pr | dem | prep | art | n | p | a | n | p | d | adj | n
- səu ju: go out of ðə insti:tju:t to zə sɑnt ənə strɑ:t | and ðen ju: gəz ði ɔðə sɑnt ənə strɑ:t to ðis to ðə sentə əv ðə taun tu ðis big rotunde
- R Gesture Unit 1
- prepar | preparation | stroke | hold | preparati | stroke | hold | preparation | stroke | hold | part retracti | h | pre | stro | preparatio | stroke
- Going out | Going o | Going alon | Goi | A roundabout

In  
Mar



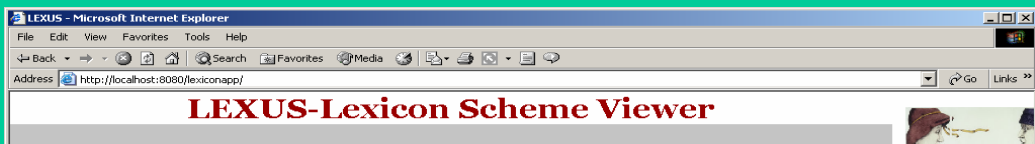
- Primary Resources:**
- Texts
  - Images
  - Sound
  - Movies



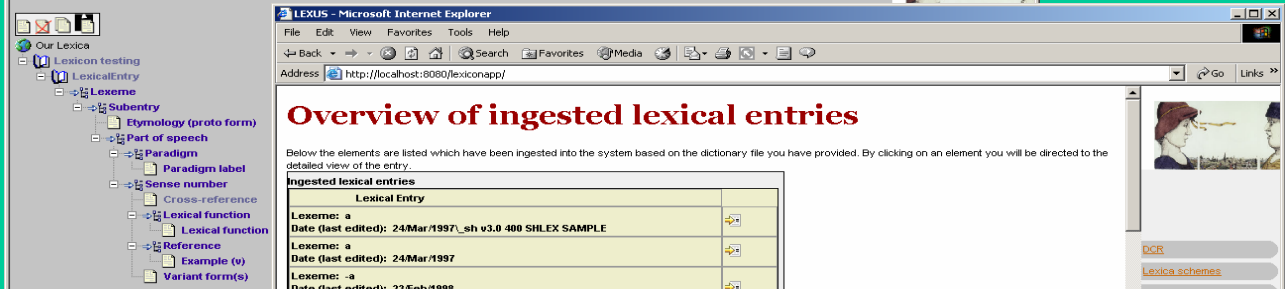
ser

REACBSWDVCPSEMNI  
 MNDOWNVTQQAQSPORC  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KND FNG FLANGUAGES  
 REACBSWDVCPSEMNI  
 MNDOWNVTQQAQSPORC  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KND FNG FLANGUAGES  
 REACBSWDVCPSEMNI

Data  
 Ingestion  
 Management



LEXUS - Microsoft Internet Explorer  
 Address: http://localhost:8080/lexiconapp/  
**LEXUS-Lexicon Scheme Viewer**

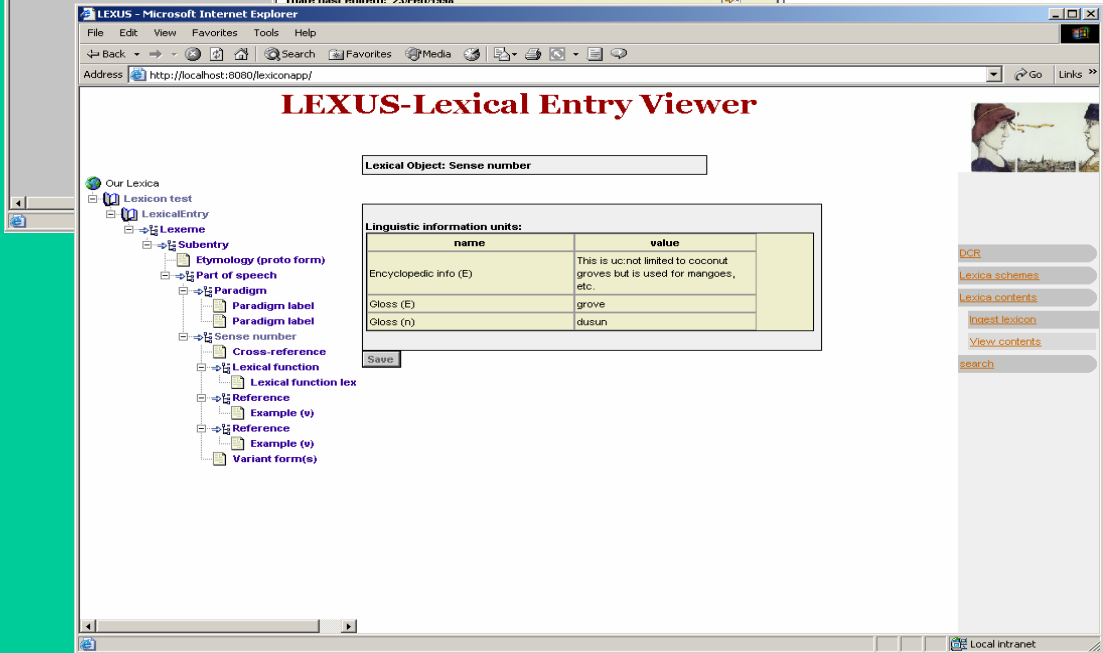


LEXUS - Microsoft Internet Explorer  
 Address: http://localhost:8080/lexiconapp/  
**Overview of ingested lexical entries**

Below the elements are listed which have been ingested into the system based on the dictionary file you have provided. By clicking on an element you will be directed to the detailed view of the entry.

**Ingested lexical entries**

Lexical Entry
Lexeme: a Date (last edited): 24/Mar/1997_sh v3.0 400 SHLEX SAMPLE
Lexeme: a Date (last edited): 24/Mar/1997
Lexeme: a Date (last edited): 23/Feb/1998



LEXUS - Microsoft Internet Explorer  
 Address: http://localhost:8080/lexiconapp/  
**LEXUS-Lexical Entry Viewer**

Lexical Object: Sense number

Linguistic information units:	
name	value
Encyclopedic info (E)	This is uc: not limited to coconut groves but is used for mangoes, etc.
Gloss (E)	grove
Gloss (m)	dusun

Save



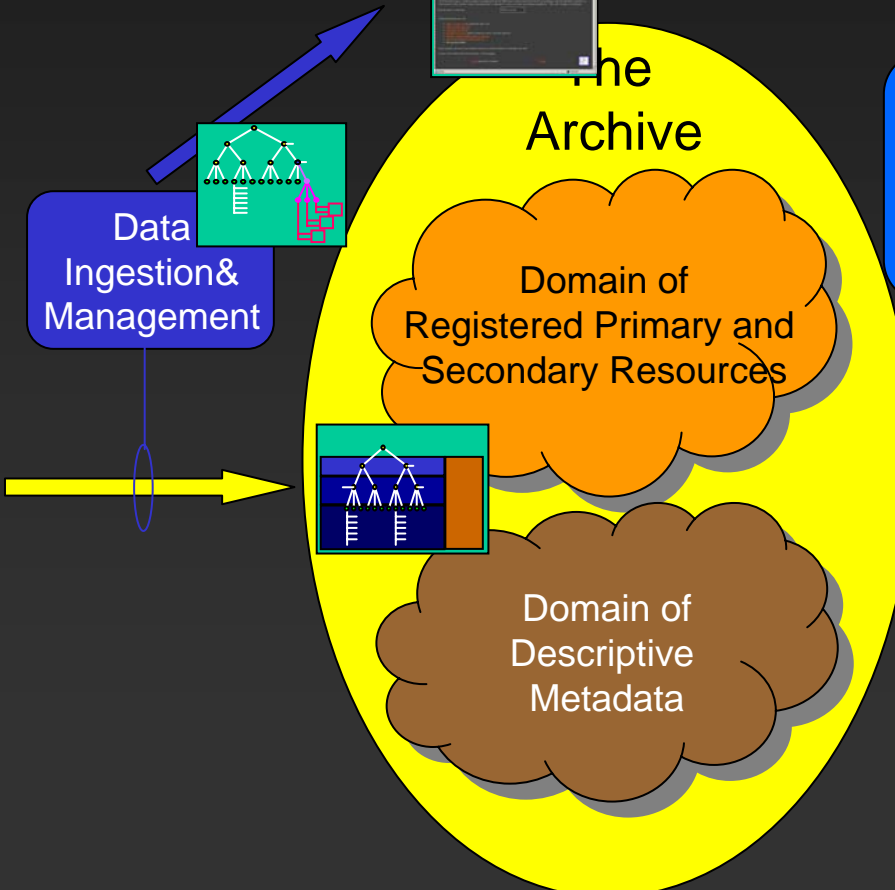
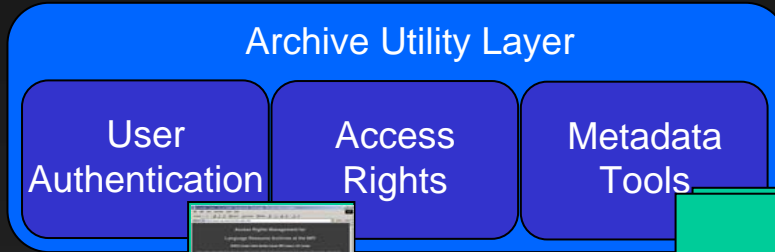
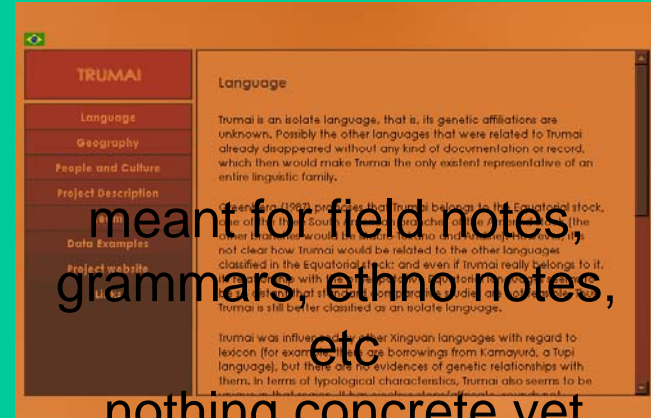
Primary Resources:  
 Texts  
 Images  
 Sound  
 Movies

- [DCR](#)
- [Lexica schemes](#)
- [Lexica contents](#)
- [Ingest lexicon](#)
- [View contents](#)
- [search](#)



ser

REACBSWDVCPSMEN  
 MDOWNVTTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDENGLANGUAGES  
 REACBSWDVCPSMEN  
 MDOWNVTTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDENGLANGUAGES  
 REACBSWDVCPSMEN

TRUMAI

Language

Trumai is an isolate language, that is, its genetic affiliations are unknown. Possibly the other languages that were related to Trumai already disappeared without any kind of documentation or record, which then would make Trumai the only extant representative of an entire linguistic family.

Project Description

Greenberg (1993) proposes that Trumai belongs to the Equatorial stock, the other members of which are now extinct. (The lower branches would be the Koro and the Mursi languages.) It is not clear how Trumai would be related to the other languages classified in the Equatorial stock; and even if Trumai really belongs to it, it is not clear with which of the other languages it is most closely related. It is not clear what stock Trumai belongs to. It is not clear if Trumai is still better classified as an isolate language.

Project website

Trumai was influenced by other Xinguan languages with regard to lexicon (for example, the borrowings from kamayurá, a Tupi language), but there are no evidences of genetic relationships with them. In terms of typological characteristics, Trumai also seems to be

meant for field notes, grammars, ethno notes, etc

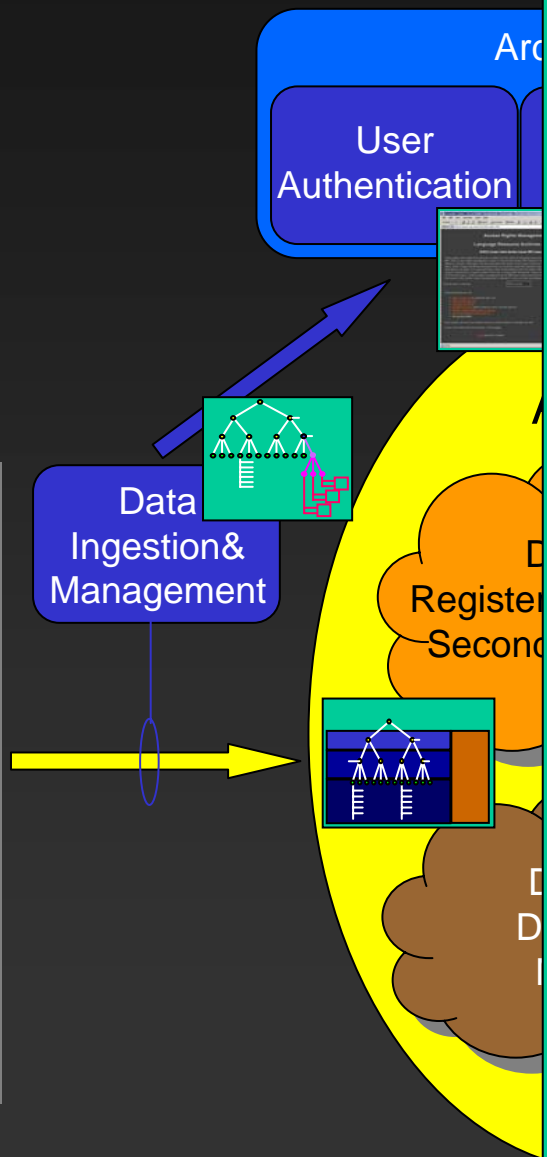
nothing concrete yet but least complex to implement



**Primary Resources:**

- Texts
- Images
- Sound
- Movies

REACBSWDVCPSEMEN  
 MNDOWNVTTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDFNGLANGUAGES  
 REACBSWDVCPSEMEN  
 MNDOWNVTTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDFNGLANGUAGES  
 REACBSWDVCPSEMEN



**Primary Resources:**  
 Texts  
 Images  
 Sound  
 Movies

**Overview of ingested lexical entries**

Lexical Entry
Lexeme: a Date (last edited): 24Mar1997_uh v3.0 466 SHELX SAMPLE
Lexeme: a Date (last edited): 24Mar1997
Lexeme: a Date (last edited): 23Feb1996
Idiosyncrasy number: 1
Lexeme: a

**LEXUS-Lexical Entry Viewer**

Lexical Object: Sense number

Encyclopedic information source:	value
Encyclopedic info (E)	This is not linked to coconut groves but is used for mangoes, etc.
Class (E)	grove
Class (N)	duvan

**TRUMAI**

Language

Trumai is an isolate language, that is, its genetic affiliations are unknown. Possibly the other languages that were related to Trumai already disappeared without any kind of documentation or record, which then would make Trumai the only extant representative of an entire linguistic family.

Greenberg (1987) proposes that Trumai belongs to the Equatorial stock, one of the three South American branches of the Amerind stock (the other branches would be Macro-Tukano and Andino). However, it is not clear how Trumai would be related to the other languages classified in the Equatorial stock; and even if Trumai really belongs to it, its relationship with the other putative Equatorial languages seems to be so distant that standard comparative studies are not feasible. Thus, Trumai is still better classified as an isolate language.

Trumai was influenced by other Xinguan languages with regard to lexicon (for example, there are borrowings from Kamayuro, a Tupi language), but there are no evidences of genetic relationships with them. In terms of typological characteristics, Trumai also seems to be

我们 先 看看 菜单。  
 Wǒmen xiān kàn kàn càidān .

Repräsentation:  
 Chinesisch | Pinyin | Deutsch | Glossen

Data Ingestion & Management



ISO 12620 data categories are listed below. Please select a data

**ISO 12620 data categories**

Data categorie	
abbreviation	A designation from a long concept.
abkhazian	the alpha-3 (terminologi
achinese	the alpha-3 (terminologi
acoli	the alpha-3 (terminologi
acronym	An abbreviat the compon or from syll syllabically.
adangme	the alpha-3 (terminologi
adjectiveClass	A categorizati whether it per of objects.
administrativeStatus	The status of assignment to certain worki
admittedTerm	A term rated s acceptability r term.
admittedTermAdmnSts	A term rated s acceptability r term.
advahe	the alpha-3 (b

mo = morpho  
n = noun

...

The screenshot shows two browser windows. The top window is Mozilla displaying an audio player for 'W-Spch' with a video thumbnail and playback controls. The bottom window is Microsoft Internet Explorer displaying the 'LEXUS-Lexical Entry Viewer' interface. The interface includes a search box for 'Lexical Object: Sense number' and a table of linguistic information units. A table is highlighted with a black box:

name	value
Encyclopedic info (E)	tree in southern Mexico, groves but is used for mangoes.
Class (E)	grove
Class (N)	Autun



**Primary Resources:**  
Texts  
Images  
Sound  
Movies

# The Problem

this is not the same for a stupid search engine

Annotation

trans	dog
POS	noun

Annotation

ortho	dog
PS	n

Lexicon

dog	form
no	wordclass
?	?

this is not the same for a stupid search engine

# Central Solution

trans	dog
POS	noun

*trans = cat 107, POS = cat 229, noun = cat 531*

ortho	dog
PS	n

*ortho = cat 107, PS = cat 229, n = cat 531*

dog	form
no	wordclass
?	?

*form = cat 107,  
wordclass = cat 229,  
no = cat 531*

Central  
ISO  
DCR

cat 107 = orthographic transcription

cat 229 = part-of-speech

cat 531 = noun

contains all relevant  
linguistic definitions  
can refer to them

given linguistic differences  
not realistic

# Individual Solution

trans  
POS

dog
noun

ortho  
PS

dog
n

dog	form
no	wordclass
?	?

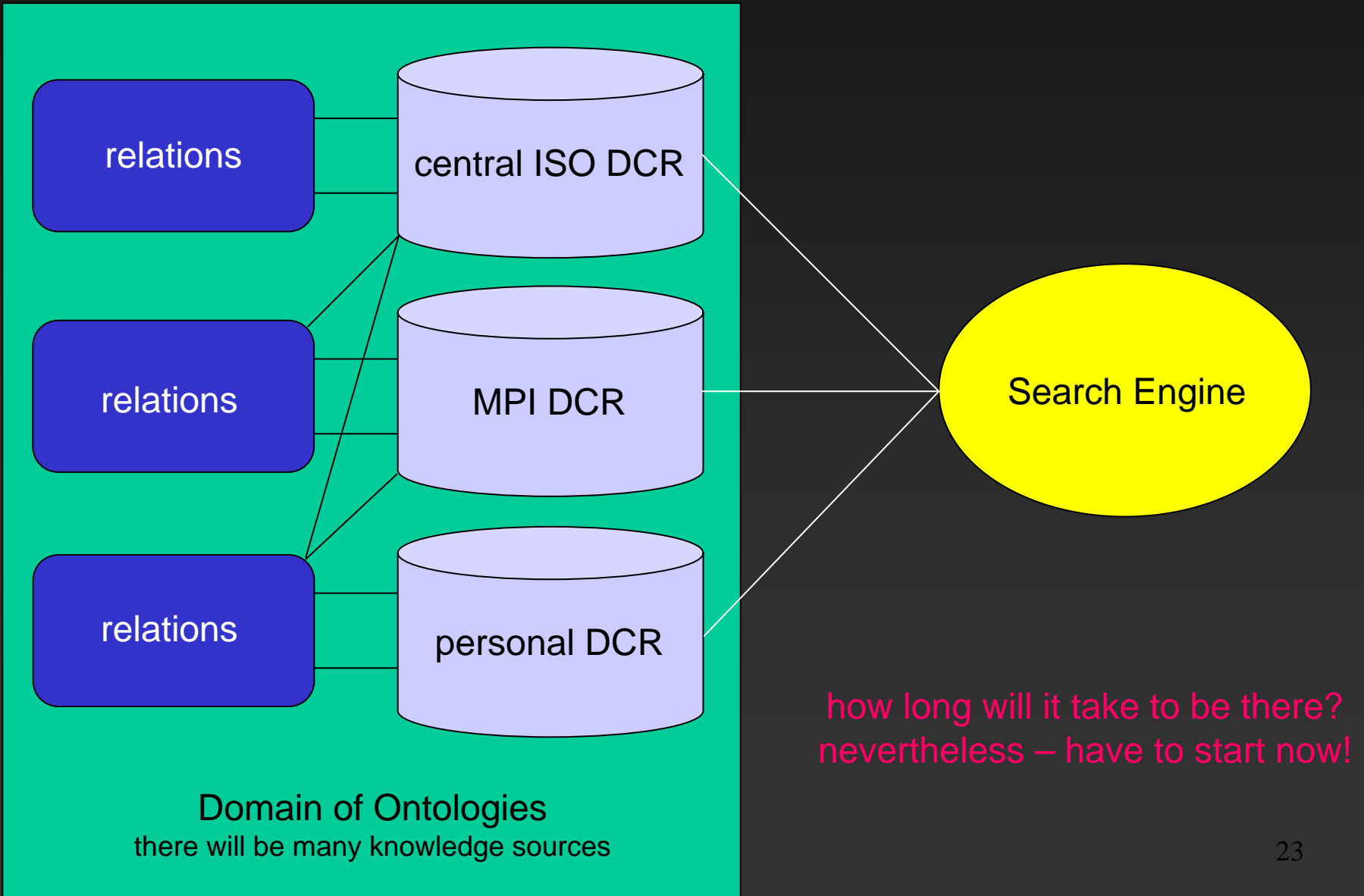
Linguist's  
mapping  
file

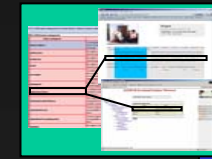
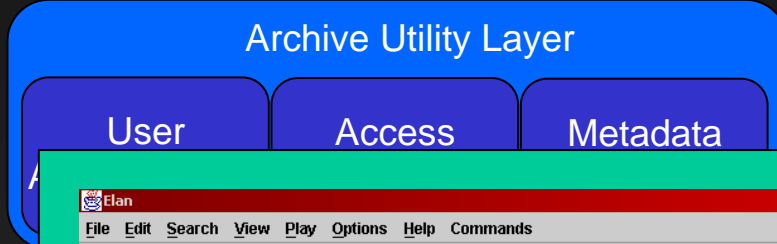
trans = ortho = form  
POS = PS = gramcat  
n = noun = no

means lot of work for all  
individuals

given time constraints not  
realistic  
will start with this version

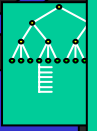
# Proper Solution





Ontological Knowledge

Data Ingestion & Management



**Elan**  
File Edit Search View Play Options Help Commands

Selection Mode Loop Mod

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

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

00:00:15.330 00:00:26.600 Selection: 15330 17450 2120

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

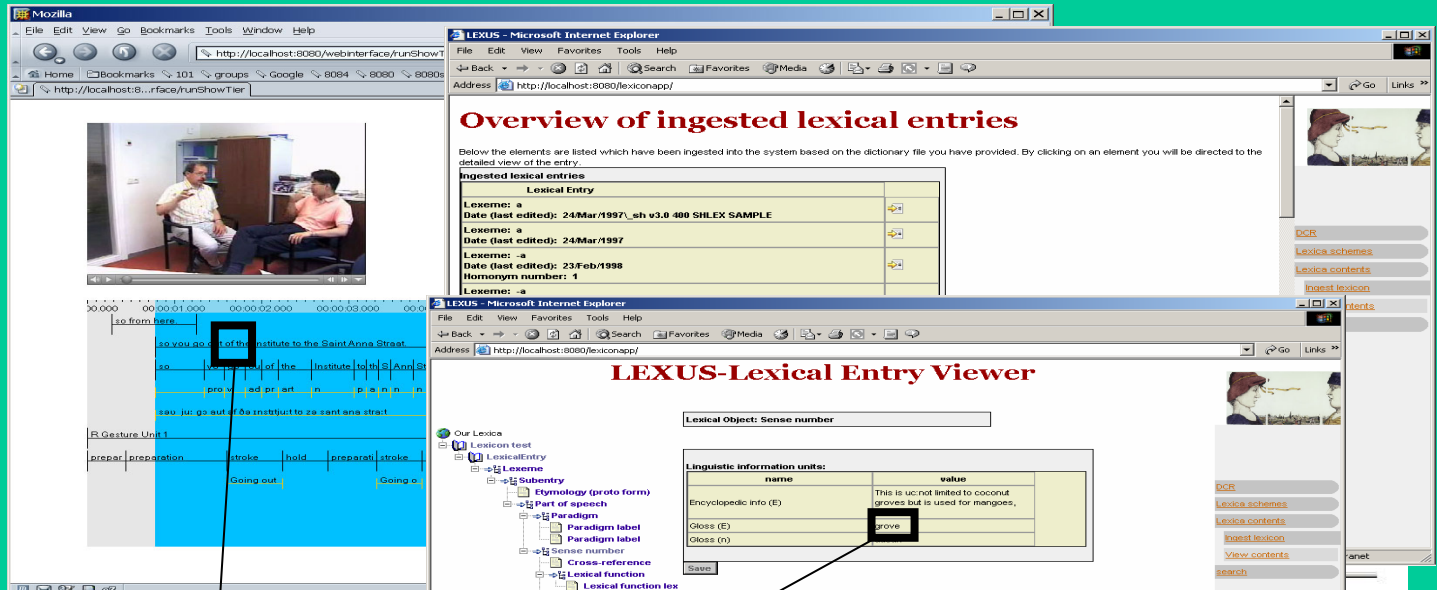
W-POS  
adv pro v art n n pro v adv pro v post prep dem n n pr v ad pr art n

W-IPA  
ðen ju ɔf ɔa ɔa sa ɔh kle ɔf ju ɔk ɔm da ɔh ju ɔn ɔa a f a s tra ja nus ple ɔn ju ɔk ɔm da ɔh t ɔ ra ɔh a v a

YET FIRST DOWNLOAD  
ANNOTATE AND UPLOAD  
ONLINE ANNOTATION  
LATER

- Primary Resources:**
- Texts
  - Images
  - Sound
  - Movies





**Overview of ingested lexical entries**

Below the elements are listed which have been ingested into the system based on the dictionary file you have provided. By clicking on an element you will be directed to the detailed view of the entry.

Lexical Entry
Lexeme: a
Date (last edited): 24/Mar/1997_sh v3.0 400 SHLEX SAMPLE
Lexeme: a
Date (last edited): 24/Mar/1997
Lexeme: a
Date (last edited): 23/Feb/1998
Homonym number: 1
Lexeme: a

**LEXUS-Lexical Entry Viewer**

Lexical Object: Sense number

Linguistic information units:	name	value
Encyclopedic info (E)		This is uc-not limited to coconut groves but is used for mangoes.
Gloss (E)		grove
Gloss (N)		

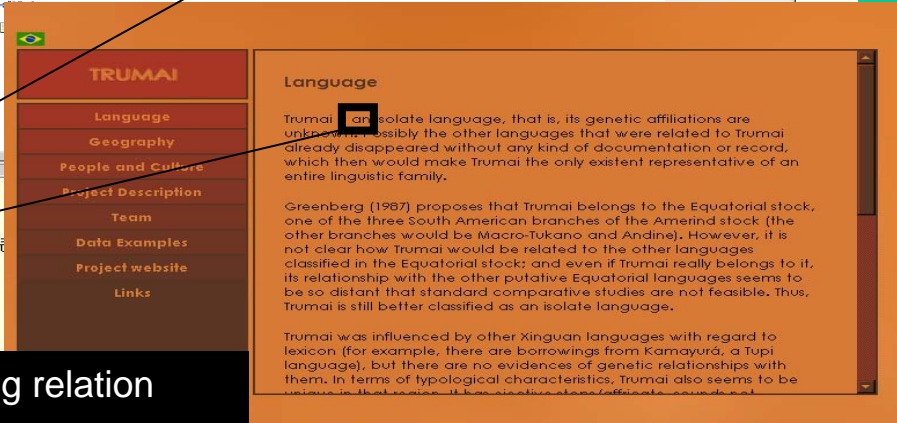
Ing  
Mar



**Primary Resources:**

- Texts
- Images
- Sound
- Movies

Comment: This is an interesting relation  
 Type: Semantic  
 Author: Peter Wittenburg  
 Date: 27.9.2004



TRUMAI	
Language	
Geography	
People and Culture	
Project Description	
Team	
Data Examples	
Project website	
Links	

**Language**

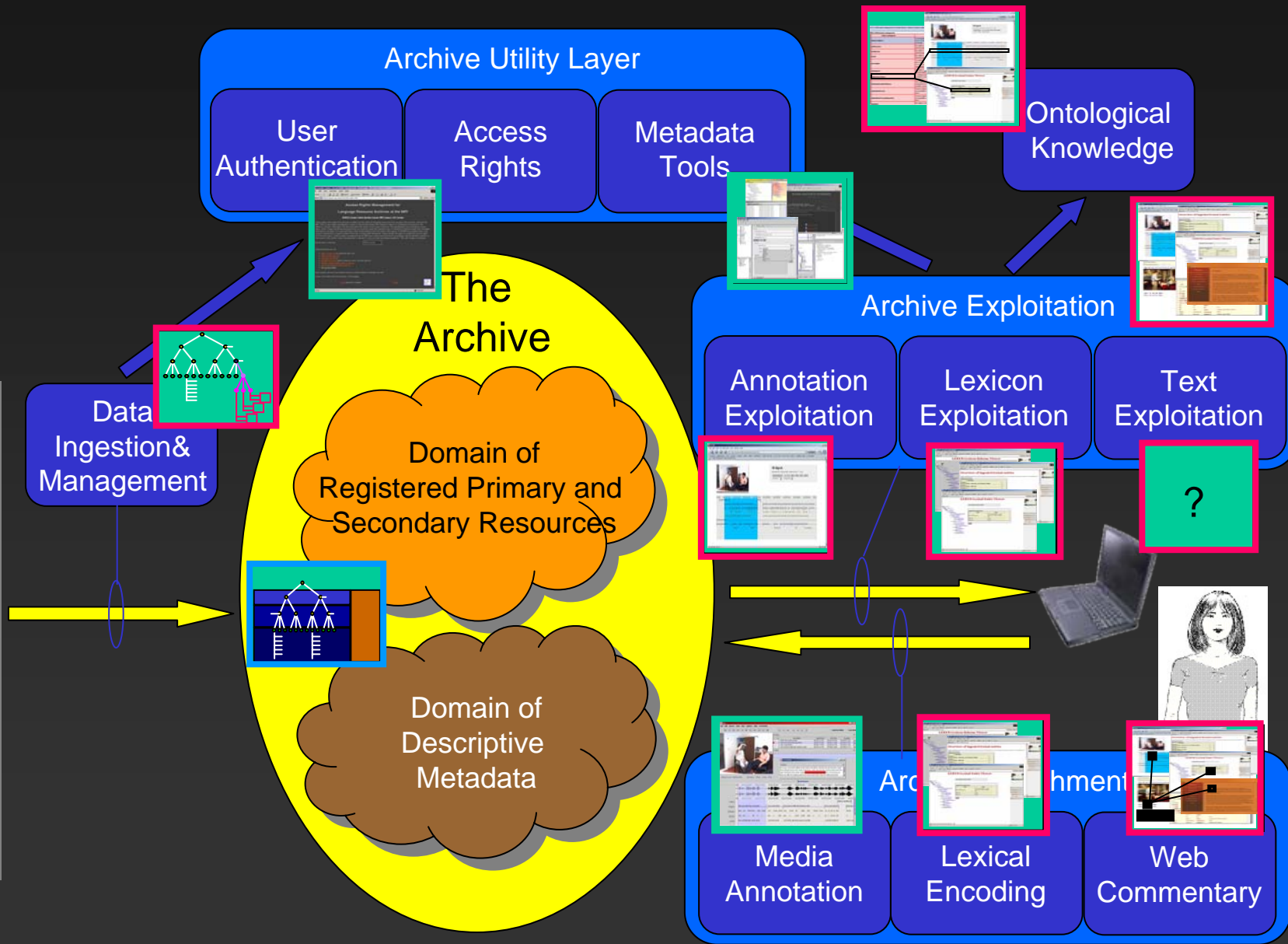
Trumai is an isolate language, that is, its genetic affiliations are unknown. Possibly the other languages that were related to Trumai already disappeared without any kind of documentation or record, which then would make Trumai the only existent representative of an entire linguistic family.

Greenberg (1987) proposes that Trumai belongs to the Equatorial stock, one of the three South American branches of the Amerind stock (The other branches would be Macro-Tukano and Andino). However, it is not clear how Trumai would be related to the other languages classified in the Equatorial stock; and even if Trumai really belongs to it, its relationship with the other putative Equatorial languages seems to be so distant that standard comparative studies are not feasible. Thus, Trumai is still better classified as an isolate language.

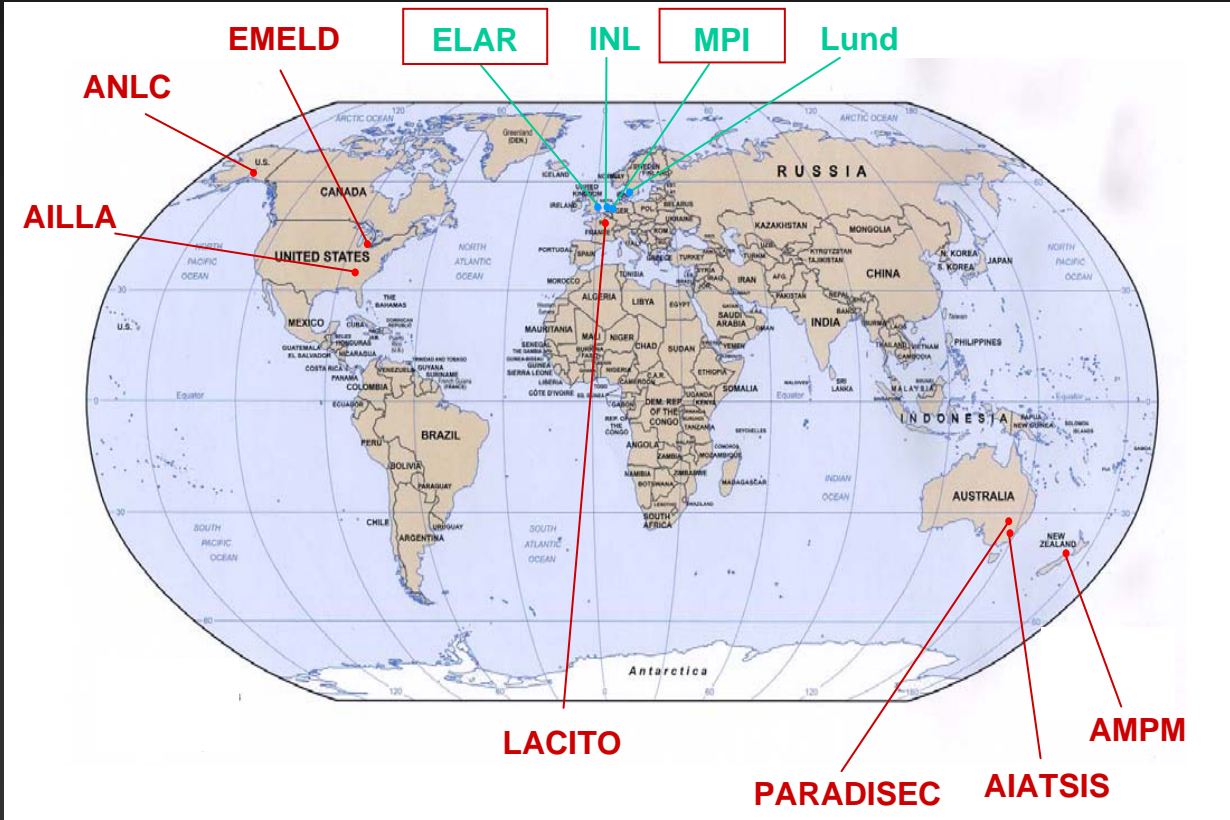
Trumai was influenced by other Xinguan languages with regard to lexicon (for example, there are borrowings from Kamayurá, a Tupi language), but there are no evidences of genetic relationships with them. In terms of typological characteristics, Trumai also seems to be unique in that respect. It has a rich class of affixes and a complex

du  
 vorher; im Voraus; zuerst  
 Herr  
 Messer und Gabel

REACBSWDVCPSMENK  
 MDOWNVTTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDFNGFLANGUAGES  
 REACBSWDVCPSMENK  
 MDOWNVTTQQAQSPOR  
 VKDUKOPDOCUMENTA  
 EISEWNSOFEENDANG  
 KNDFNGFLANGUAGES  
 REACBSWDVCPSMENK

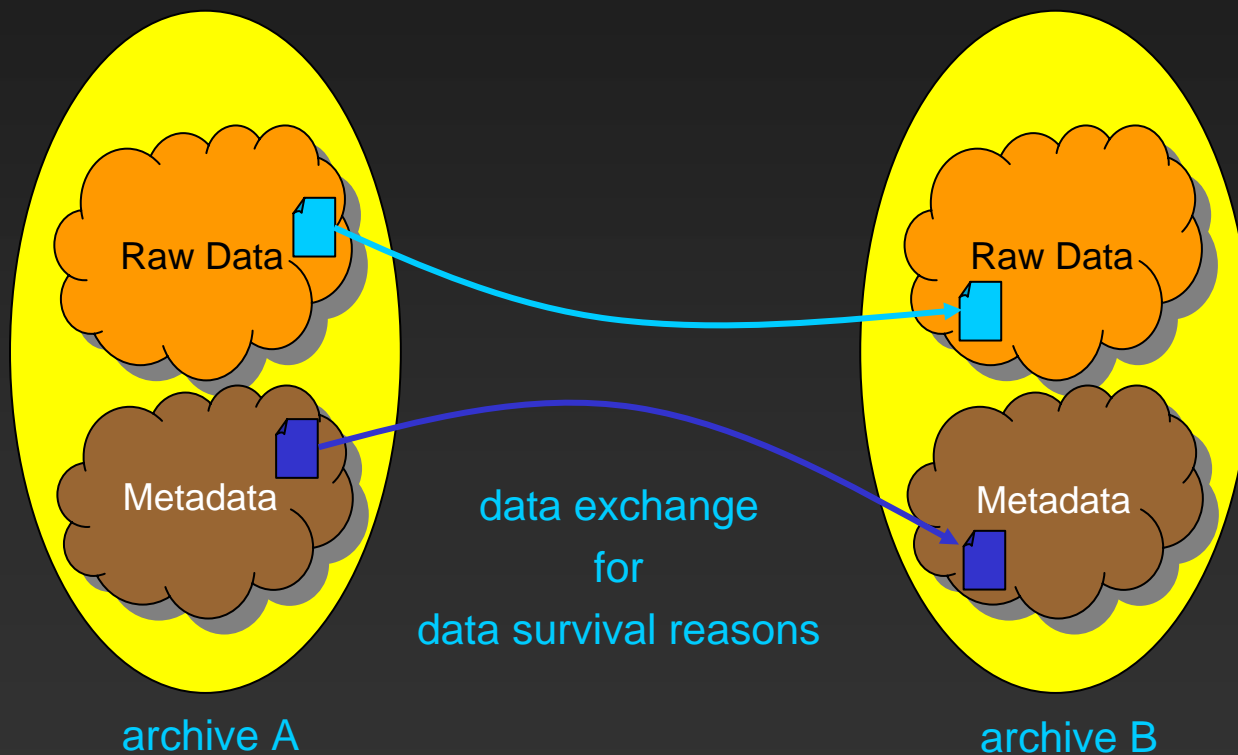


# Cross-Archive Dimension DELAMAN / DAM-LR Visions

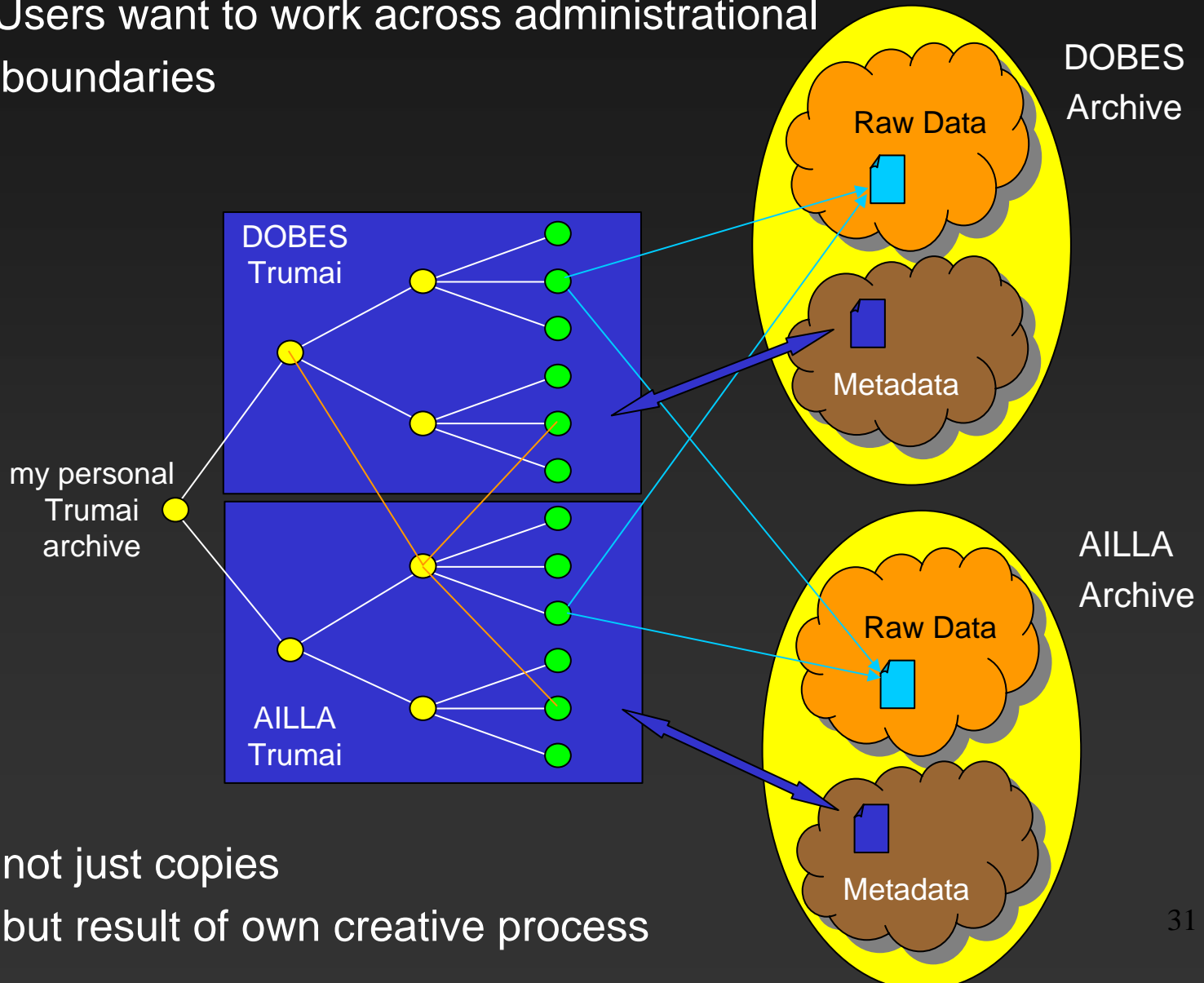




- have to take care of long-term data preservation
- only chance is world-wide distribution



- Users want to work across administrative boundaries



not just copies  
but result of own creative process

- it's about future usage scenarios with distributed archives
- it's about federated language resource archives
- it's about eScience scenarios in linguistics
- want to exchange data automatically (list driven)
- want to allow people to create integrated virtual working spaces
- want to have an integrated access management domain (one identity, rights go with the copies, ...)
- first talks in Nijmegen and at HRELP workshops 2003
- foundation at PARADISEC meeting in Sydney 2003
- last workshop in Nijmegen November 2004
  - linguists
  - archivists
  - (GRID) technologists



- much technology to achieve our goals is available
  - **A-Select** authentication system
  - **Shibboleth** authorization system
  - **Handle System** for URID resolving
  - Distributed metadata environment such as **IMDI**
  - **Storage Request Broker** for federated resources
  - **Web-Services** for layered services
  - ...

DELAMAN Web-Site

[www.delaman.org](http://www.delaman.org)

DELAMAN Workshop-Site

[www.mpi.nl/delaman/workshop](http://www.mpi.nl/delaman/workshop)

DOBES Web-Site

[www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)

MPI Archive Web-Site

[www.mpi.nl/world/corpus](http://www.mpi.nl/world/corpus)

MPI Tools Web-Site

[www.mpi.nl/tools](http://www.mpi.nl/tools)