



# Eureka! User friendly access to the MPI linguistic data archive

Alexander Koenig

Jacqueline 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

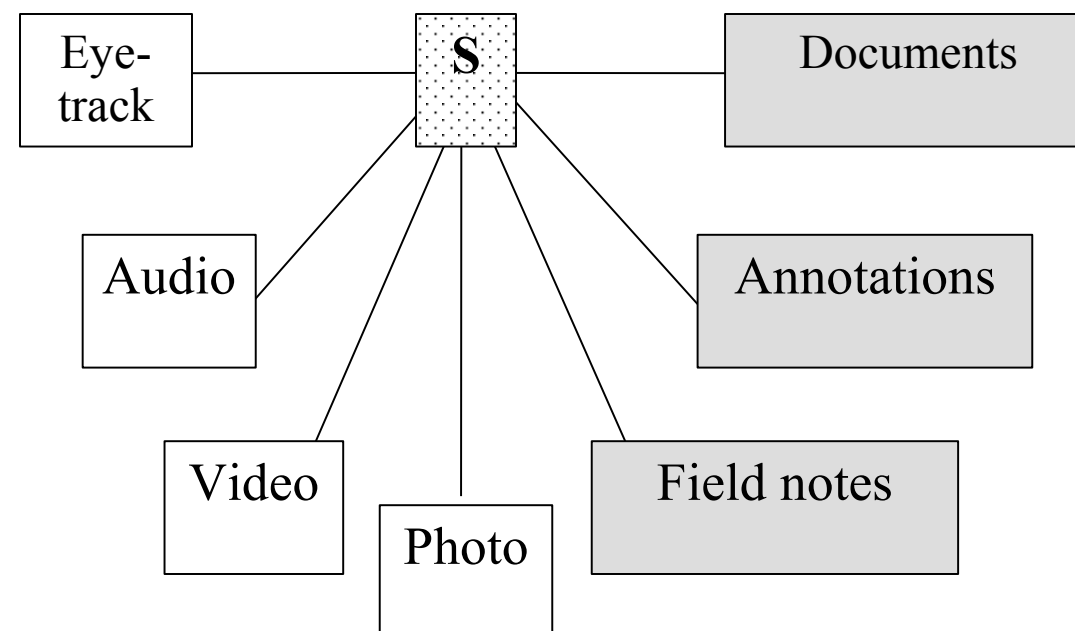
DoBeS projects

Corpora from other research institutions

- Corpus Nederlandse Gebarentaal
- European Science Foundation's Second Language Acquisition by Adult Immigrants
- Dutch Bilingual Database



- 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)

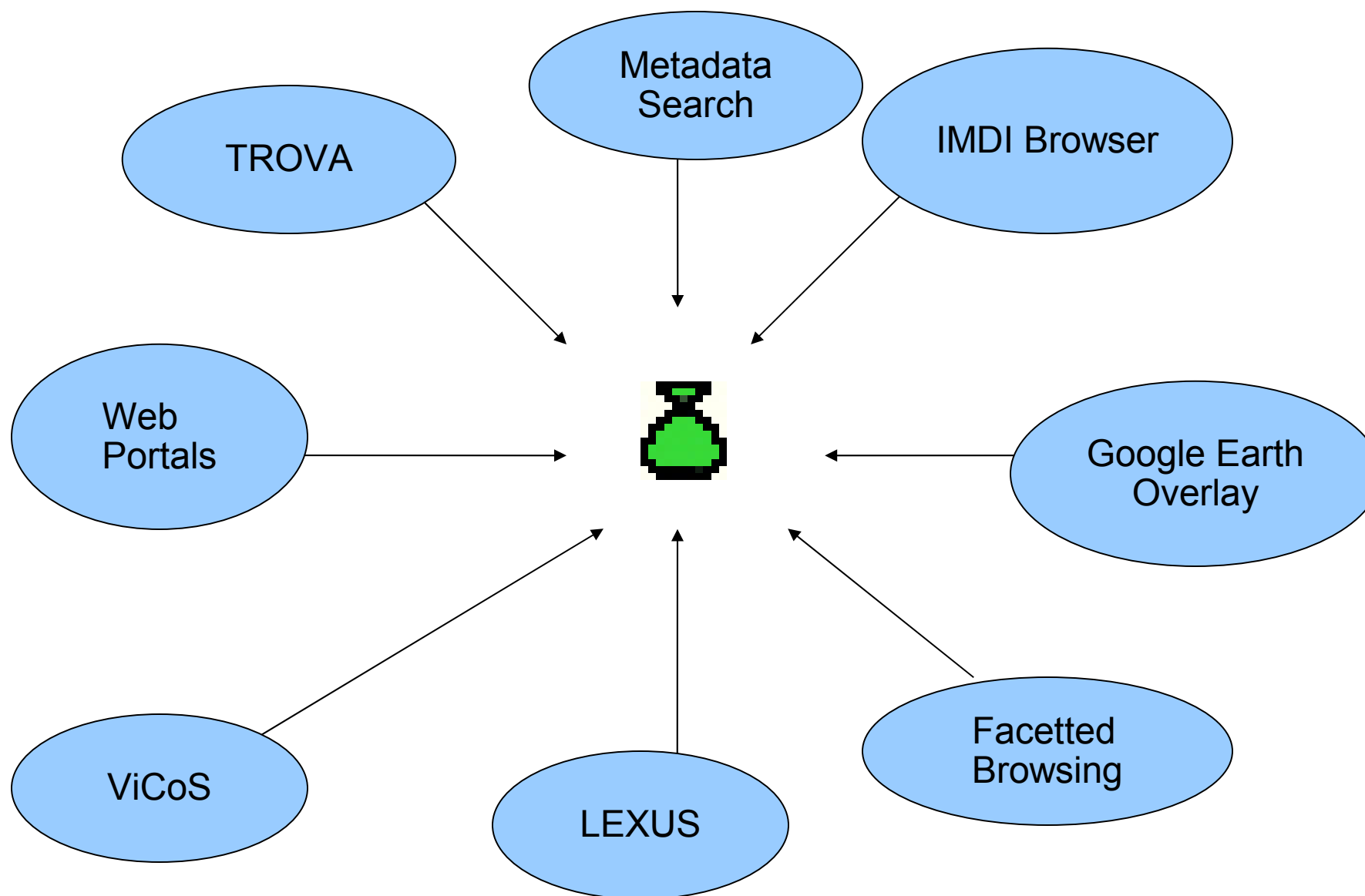




- every resource and metadata description has a URID (<http://www.handle.net/>)
- all data in principle available online
  - <http://corpus1.mpi.nl>
- complex access management system regulates who can access what



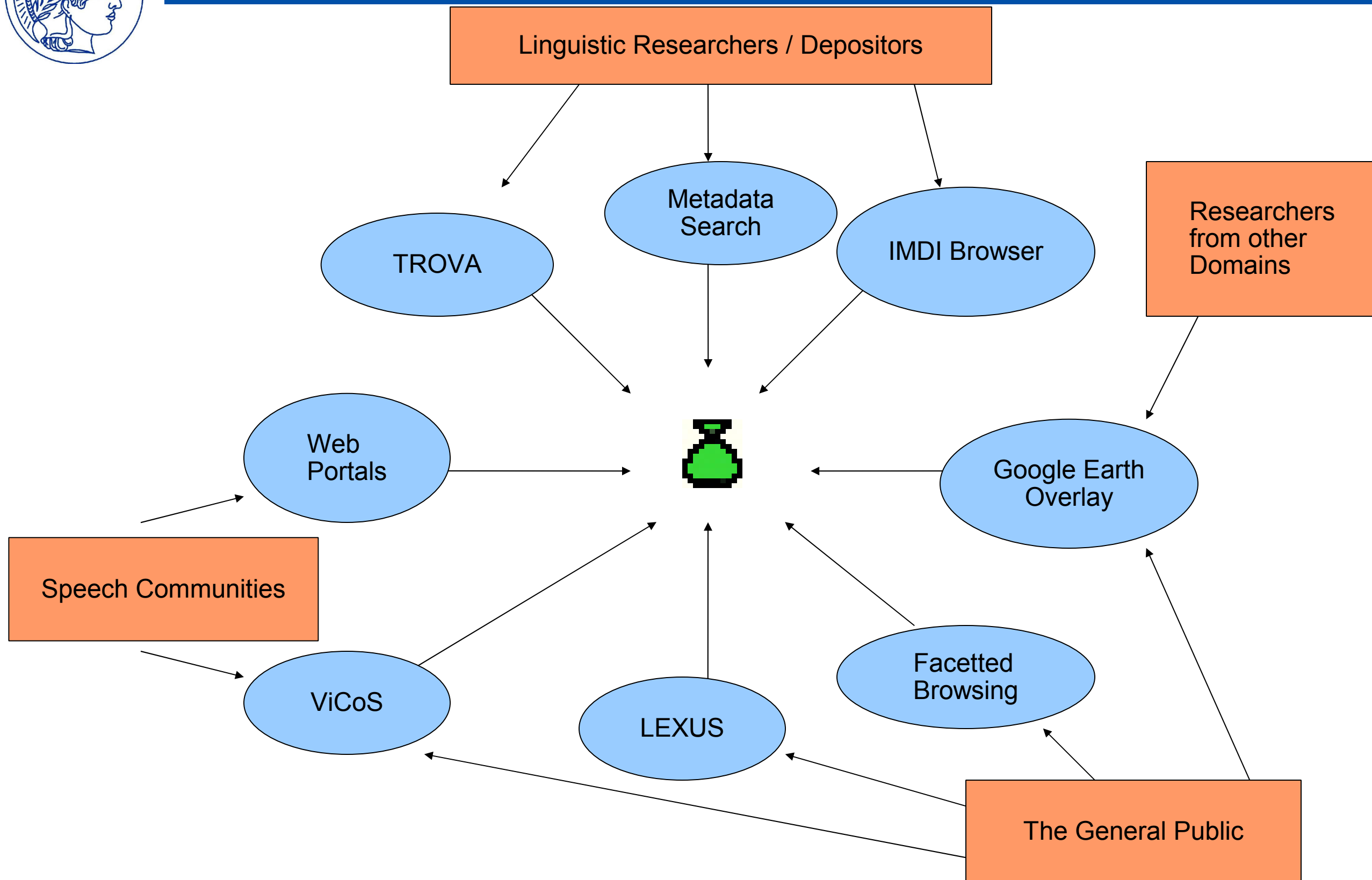
# Accessing the Data







# Accessing the Data



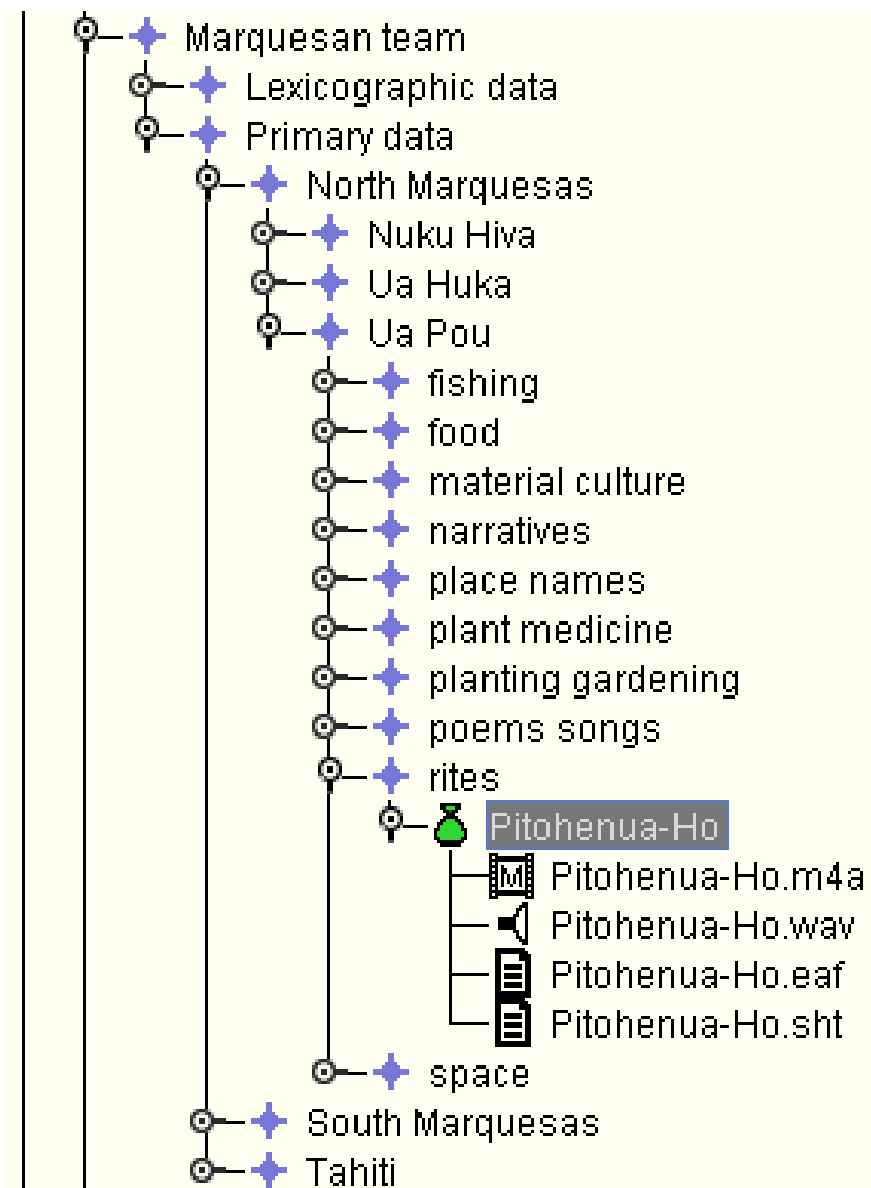




# **The Traditional Method – The IMDI Browser**



- 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

Navigation tree showing a hierarchical structure of data:

- D
- BeS archive
  - WelcomeToDOBES.html
  - introduction.html
- Aweti
- Bakairi, Katxuyana and Mawe project
- Baure
- Beaver Archive
- Cashinahua project
- Chaco languages
- Chintang and Puma Documentation Project
- Chipaya project
- Chontal
- ECLING project
- Enets and Forest Nenets
- Gente del Centro – People of the Center
- Gorani
- Hocank team
- Iwaidja team
- Jaminjungan and Eastern Ngumpin
- Kola Saami Documentation Project
- Kuikuro Team
- Kurumba
- Lacandon Cultural Heritage
- Marquesan team
  - Lexicographic data
  - Primary data
    - North Marquesas
      - Nuku Hiva
      - Ua Huka
      - Ua Pou
        - fishing
        - food
        - material culture
        - narratives
        - place names
        - plant medicine
        - planting gardening
        - poems songs
        - rites
          - Pitohenua-Ho
            - Pitohenua-Ho.m4a
            - Pitohenua-Ho.wav
            - Pitohenua-Ho.eaf
            - Pitohenua-Ho.sht

Session details for Pitohenua-Ho:

<b>Session</b>
<b>Name</b> Pitohenua-Ho
<b>Title</b> Short explanation of after bir
<b>Date</b> Unspecified
<b>Description</b>
This text explain briefly some birth rites, namely
<b>Location</b>
<b>Project</b> Marquesan-DOBES
<b>Keys</b>
<b>text</b> explanation
<b>content</b> birth rites
<b>language</b> Ua Pou dialect
<b>Content</b>
<b>Actors</b>
<b>Actor</b> GC
<b>Actor</b> Ho
<b>MediaFile</b>
<b>MediaFile</b>
<b>WrittenResource</b>
<b>WrittenResource</b>
<b>Anonyms</b>
<b>References</b>



# The Traditional Method – The IMDI Browser

- Resources can be viewed using ANNEX

The screenshot displays the ANNEX web interface. At the top, there is a navigation bar with a red ribbon icon, the text "ANNEX", and links for "manual", "? settings", and "user: anonymous login".

The main content area is divided into several panels:

- Text:** A sidebar menu with options: Text, Grid, Subtitle, Waveform, Timeline (highlighted), and Combined.
- Video display:** A video player showing two men in a room. Below the video are playback controls (play, stop, previous, next) and a "Settings" button.
- Media information:** A panel displaying:
  - Resource: elan-example1.eaf
  - Media file: elan-example1.mp4
  - Elapsed time: 00:00:00:000
  - Selected chunk:
    - Begin time: -
    - End time: -
    - Text: -
- Mini Data Frame:** A panel with a dropdown menu for "Tier" (set to "none") and a "Font size" dropdown (set to "14").

Below these panels is a large **Timeline** section. It features a horizontal time axis from 00:00:00 to 00:00:05:50. The timeline is organized into tracks for different data types:

- K-Spch:** Shows speech segments with text like "so from here." and "yeah".
- W-Spch:** Shows word-level speech segments with text like "so you go out of the Institute to the Saint Anna Straat." and "and then you go the other. Sai".
- W-Words:** Shows individual words from the speech segments.
- W-POS:** Shows part-of-speech tags for the words.
- W-IPA:** Shows International Phonetic Alphabet (IPA) transcriptions for the words.
- W-RGU:** Shows Gesture Units, with a tooltip for "R Gesture Unit 1" displaying:
  - Begin time: 00:00:00:780
  - End time: 00:00:04:090
  - Text: sau ju: go aut af ðə insttju:t to zə sant ana strɑ:t
- W-RGph:** Shows gesture phases like "preparation", "stroke", "hold", and "preparation".

On the left side of the timeline, there are navigation buttons: "Play selection", "Clear selection", and a set of directional arrows (|<, >|, <<, >>, <, >, +, -). Below these are playback options: "Play screen by screen" (selected) and "Play continually". At the bottom left, there is a "Tier text font:" dropdown menu set to "Arial Unicode MS".



**More than just Browsing  
—  
Metadata and Content Search**



- Metadata is always open and can be searched

**Metadata search**  Show occurrences

within 1 selected corpus: DoBeS archive [12570 sessions]

**Key word search** **Standard search** **Advanced search**

Session	Content	Languages	Language	(X)	Name	Chontal
Session	Content	Genre				Discourse

77 matches within 12570 selected sessions found.

/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/RSDiablo  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/RSpobre  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/ael1reared  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/ael7vida  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/aer1panka  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/aer2infiel  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/aerVida1of2  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/aerVida2of2  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/apom1vida  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Personal stories - Historias personales/iz1biida  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Towns - Pueblos/Rio Seco/RSlafelay  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Towns - Pueblos/San Felipe/apm2sanfelipe  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Towns - Pueblos/San Miguel del Puerto/aer8shooting  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Towns - Pueblos/San Pedro Huamelula/ael5diluvio  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/01 History - Historia/Towns - Pueblos/Santiago Astata/ael2astata  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/02 Society - Sociedad/Courting - Enamorando/rg3enams1  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/02 Society - Sociedad/Courting - Enamorando/rg5enams2  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/02 Society - Sociedad/Daily life - Vida cotidiana/rg1ajutl  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/02 Society - Sociedad/Food - Comida/Kitchen  
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/02 Society - Sociedad/Food - Comida/Kitchen2



# More than just Browsing – Metadata and Content Search

- If you have access to the resources, you can also search the annotations

**Domain:** MPI CGN

EAF (12767)

History:

Mode:

<input type="text"/>	<input type="text" value="fiets"/>	<input type="text"/>	in	<input type="text" value="Tier Type: Words"/>
<input type="text" value="Fully aligned"/>	<input type="text" value="Fully aligned"/>	<input type="text"/>		
<input type="text"/>	<input type="text" value="WW"/>	<input type="text"/>	in	<input type="text" value="Tier Type: PoS"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	in	<input type="text" value="All Tiers"/>

#hits : 519  
#annotations with a hit : 519  
#annotations investigated : 4240130

Progress

<input type="text" value="*"/>	<input type="text" value="fietsen"/>	<input type="text" value="*"/>
<input type="text" value="*"/>	<input type="text" value="WW(inf,vrij,zonder)"/>	<input type="text" value="*"/>





# **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)

http://ems06.mpi.nl/cgi-bin/flamenco.cgi/flama/Flamenco

Powered by Flamenco

Save Search History and Settings Return to Search New Search Logout

search

Username WikiSysop Password ..... Log In  
[Create a New Account](#)

Show tooltip previews of subcategories

### CORPUS

<a href="#">ailla</a> (1794)	<a href="#">dobes</a> (17498)
<a href="#">andes</a> (231)	<a href="#">edo</a> (245)
<a href="#">bas</a> (7417)	<a href="#">endangeredLanguages</a> (17941)
<a href="#">bifo</a> (1178)	<a href="#">esf</a> (2854)
<a href="#">cgn</a> (92029)	<a href="#">grtp</a> (34)
<a href="#">coralRom</a> (772)	<a href="#">more...</a>
<a href="#">dbd</a> (2122)	

### CONTINENT

<a href="#">Africa</a> (3763)	<a href="#">North-America</a> (7620)
<a href="#">Asia</a> (13410)	<a href="#">Oceania</a> (4822)
<a href="#">Australia</a> (4822)	<a href="#">South-America</a> (14496)
<a href="#">Europe</a> (134516)	<a href="#">Unknown</a> (31)
<a href="#">Middle-America</a> (1737)	

### COUNTRY

<a href="#">Argentina</a> (816)	<a href="#">China</a> (457)
<a href="#">Australia</a> (4839)	<a href="#">Colombia</a> (835)
<a href="#">Belgium</a> (29444)	<a href="#">Costa Rica</a> (24)
<a href="#">Bolivia</a> (5611)	<a href="#">Czech Republic</a> (1600)
<a href="#">Botswana</a> (304)	<a href="#">East Timor</a> (1152)
<a href="#">Brazil</a> (3435)	<a href="#">more...</a>
<a href="#">Canada</a> (3417)	

### LANGUAGE

<a href="#">!Xoon</a> (348)	<a href="#">Akurio</a> (4)
<a href="#">!Xu</a> (2)	<a href="#">Amurdak</a> (12)
<a href="#">'Njohan</a> (412)	<a href="#">Andoke</a> (18)
<a href="#">=/Hoan</a> (2)	<a href="#">Arabic</a> (18)
<a href="#">Achuar</a> (10)	<a href="#">Arabic, Moroccan Spoken</a> (481)
<a href="#">Afrikaans</a> (778)	<a href="#">more...</a>
<a href="#">Akkala Saami</a> (28)	

### ORGANISATION

[Académie marquisienne \(Tuhuna 'Eo](#) [CERTEC / Humanistlaboratoriet](#) (3)

### GENRE

<a href="#">Analysis</a> (4)	<a href="#">Conversation</a> (582)
<a href="#">Book</a> (4)	<a href="#">cultural activity</a> (4)
<a href="#">Cantos (religiosos)</a> (6)	<a href="#">Cultural data</a> (234)
<a href="#">Cantos religiosos</a> (38)	<a href="#">Culture</a> (320)
<a href="#">Cantos rituales</a> (38)	<a href="#">Dance</a> (4)
<a href="#">commands</a> (22)	<a href="#">more...</a>
<a href="#">Concert</a> (4)	

### INTERACTIVITY

<a href="#">interactive</a> (68363)	<a href="#">semi-interactive</a> (22530)
<a href="#">non-interactive</a> (74994)	<a href="#">unknown</a> (64)

### PLANNING TYPE

<a href="#">performer-planned</a> (4)	<a href="#">spontaneous</a> (48222)
<a href="#">planned</a> (23788)	<a href="#">unknown</a> (24)
<a href="#">semi-spontaneous</a> (87246)	

### INVOLVEMENT

<a href="#">elicited</a> (33542)	<a href="#">non-elicited</a> (118804)
<a href="#">no-observer</a> (7553)	<a href="#">unknown</a> (109)

### SOCIAL CONTEXT

<a href="#">Controlled environment</a> (8741)	<a href="#">Public</a> (60972)
<a href="#">Family</a> (4884)	<a href="#">Unknown</a> (65)
<a href="#">Private</a> (56648)	



**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

**Yel' Dnye**

**Pioneers of Island Melanesia**

Yel' Dnye (also known as Rossel, Yela, Yele, Yeljong, Yelelye) is a Papuan language spoken on Rossel Island, Louisiade Archipelago, Papua New Guinea. Although surrounded by Austronesian languages, Yel' Dnye shows little evidence of influence by them, making this language an isolate. Yel' Dnye is known as the language with the world's most complex phonemic inventory.

Project leader is [Stephen Levinson](#).

[Browseable corpus](#)  
[Language and Grammar Yel' Dnye website](#)

**LEXUS example**  
For this demonstration of Lexus use "demo" as username and "demo" as password.  
[ANNEX example 1](#) [ANNEX example 2](#) [ANNEX example 3](#)  
[Pioneers of Island Melanesia website](#)

 Village on Rossel Island

 LEXUS and ANNEX examples





# **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



# Back to the Communities – Community Web Portals

Dane-zaa Community Portal

## 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.

Stories	Learn about ...	Materials	The Archive
Personal history	Drum	Movies	Searching
Traditional stories	Food and cooking	Clickables	Navigating
Animals	Handgames	Dictionary	Download
Place	Horses	Phrasebook	Tours
	Moccassins	Calendar	Google Earth
	Moosehide	Alphabet	Studies
	Preparing meat	Posters	



## **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





# Word for word navigation - LEXUS

demo rossel lexicon

Lexical Entry View   Lexical entry   Search

List   Tree   Selections

StartLetter:  Lexicon:

Selection:

List items

- chili** one of two sides of a traditional fishing net tree sp.
- chimi chapì** hunting for shellfish on reet
- chìmo** fish type stone axe head, used as valuable
- ch:oo** fish sp. (Orangespine unicornfish, *Naso literatus*) hitching pole
- chu** bird sp, Common Dollarbird (*Eurystomus orientalis*)
- chuu** (edible) fruit tree batten (for house roof) bird sp.(see chu)
- chuu kigha** fish sp. (Coral rabbitfish, *Siganus corallinus?*) tree sp
- d:aa** fat inside a turtle fish sp, scorpion fish sp. ?*Pterois volitans*
- dââtp:ee** bird sp. owl (*Ninox themacha rosseliana*)
- dada** fish sp (black spot sea perch) more than should (tentative)
- dada** fish sp. (grunter) tree sp (qv dada y)
- dêê dmi** fins of fish
- d:êê vyono** fish sp. (an unidentified wrasse, *Cheilinus* sp.)

Remove   New

Result 22 of 348

First   Previous   Page 1   Next   Last


---

## dada

(N)

---

### fish sp (black spot sea perch)



more than

**Pintyó p:uu y:a dada doo ya,**  
*more than 10 people were with Pintyó*

**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



# Browsing Concepts - ViCoS

Back Forward Reload Stop Home [http://lux07.mpi.nl/mpi/vicos/ViCoS\\_Browser.html](http://lux07.mpi.nl/mpi/vicos/ViCoS_Browser.html)

Most Visited Getting Started Latest Headlines (Untitled) Language Archiving T... LAMUS - Language Ar... IMDI Browser Welcome to the Max ...

ViCoS - Visualising Conceptual Spaces ViCoS Editor and Navigator <http://lux07.mpi.nl...ebug/FlexViCoS.html>

**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
overview	refetch
save	colour

**Relation Types:**



## Conclusion

- 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



