



The Language Archiving Technology Domain

Alexander König

Jacqueline Ringersma

Paul Trilsbeek

Max Planck Institute for Psycholinguistics



- only 10 years since we started
- 300,000 audio, video & text resources
- 100,000 metadata descriptions
- 19 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



The Three Pillars of Archive Management

The MPI Linguistic Archive

LAMUS

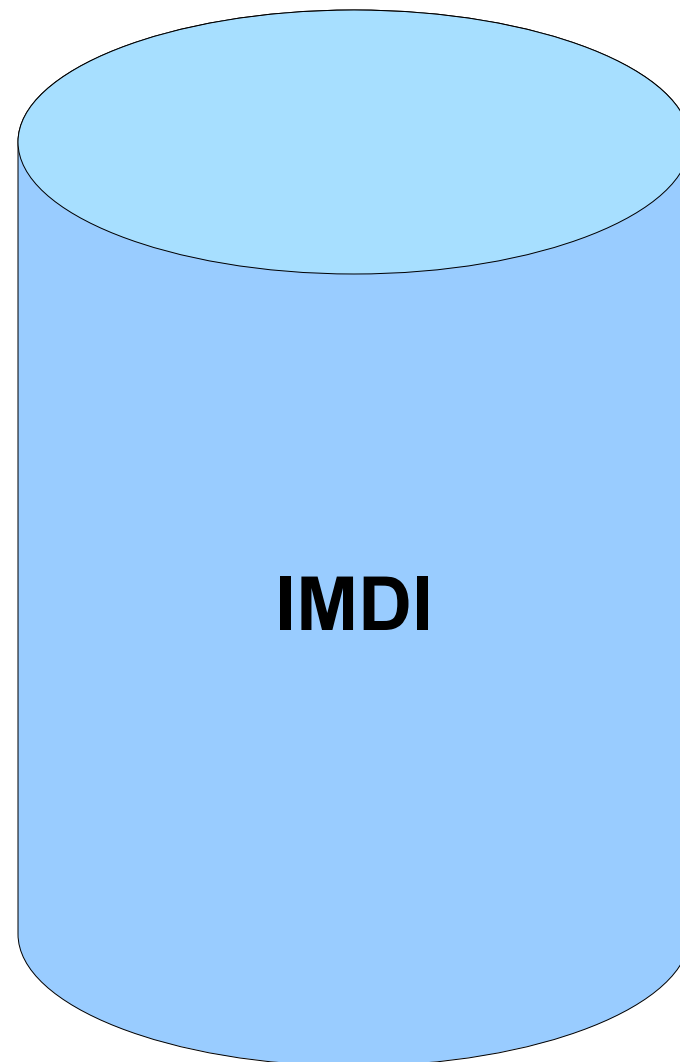
IMDI

AMS



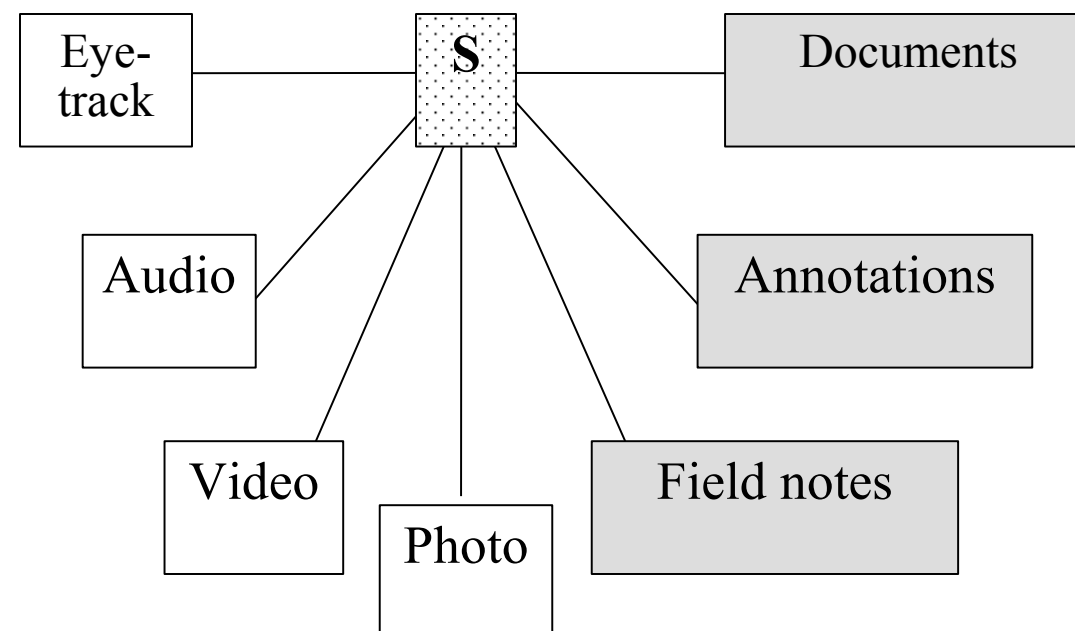
The Three Pillars of Archive Management

The ISLE Metadata Initiative Format





- all data described using (XML-based) IMDI metadata format
- every resource/md description has a URID
- central organisational concept: the session resource bundle
 - obligatory metadata
 - a set of annotations & transcriptions
 - a set of media data (video, audio, photo)





- all data in principle available online
 - <http://corpus1.mpi.nl>

IMDI Browser - Konqueror

Location: http://corpus1.mpi.nl/ds/imdi_browser/

IMDI-Browser about user: anonymous login logout

IMDI Metadata Domain

The IMDI Metadata Domain allows you to browse and search in the whole domain of linked IMDI metadata descriptions as they are registered at the IMDI portal at the MPI for Psycholinguistics. All Metadata descriptions are openly accessible, for many resources however one needs to ask access permission. Read further to find out how this is done.

Please note that many corpora are hosted at other resource centers. If a remote server is down, those resources cannot be accessed even though they are still listed within this portal.

How to explore the archive:

On the left hand side, you see the whole hierarchy of linked IMDI corpora, which you can explore by clicking on the small circles or by double-clicking the names. If you click on an item with the **right mouse button** (or control-click with a single button Mac mouse), you will get a contextual menu with additional functionality, depending on what kind of item you have selected:

- **view node:** shows the content of a file. For Metadata files, you will get the content of the metadata description. For media files, you will get to see the actual resource. For EAF, Shoebox/Toolbox and CHAT annotation files, the ANNEX annotation viewer will be launched. If access to a resource is restricted, you will get an authentication window in which you need to type your user name and

Page loaded.



- can be created/modified with IMDI-Editor

The screenshot shows the 'IMDI Metadata Editor' window. The interface is divided into several sections:

- Menu and Toolbar:** 'File View Options Help' menu and icons for file operations.
- Left Panel (Tree View):**
 - Top section: 'Standard (Session) Resource Bundle' with sub-items: Project, Content (selected), Actors, Resources, References.
 - Bottom section: 'Local Repository' with sub-items: Projects, Content, Actors, Languages, Access, Local_Repository.
- Main Editor Area:**
 - Buttons: Editor, HTML, Links.
 - Content Summary Content Information:** Genre *Unspecified*, Subgenre *Unspecified*, Task, Modalities, Social Context *Unspecified*, Event Structure *Unspecified*, Channel *Unspecified*, Subject.
 - Content Type Descriptions Languages Keys:**
 - Genre: Unspecified (CV)
 - Subgenre: Unspecified (CV*)
 - Task: (CV)
 - Modalities: (CV*)
 - Subject: (CV*)
 - Communication Context:**
 - Interactivity: Unspecified (CV)
 - Planning Type: Unspecified (CV)
 - Involvement: Unspecified (CV)
 - Social Context: Unspecified (CV)
 - Event Structure: Unspecified (CV)
 - Channel: Unspecified (CV)
 - Button: Clear Content



- Soon Arbil will be released
(testing version already available)

The screenshot displays the Arbil software interface. On the left, there is a tree view showing a hierarchy of corpora. The 'Local Corpus' section is expanded, showing a tree structure with nodes like 'Kleve-route', 'Actors', 'Peter', 'Languages', 'Peter Wittenburg', 'Standard Actor', and 'MediaFiles'. The 'Testing' section is also visible, containing files like 'AMS', 'docbook.dtd', 'elan-example1.eaf', 'elan-example1.pfs', 'Lamus', 'manualXML', 'oriola.eaf', and 'oriola.pfs'.

On the right side, there are two data tables. The top table, titled 'IMDI Field', shows the following data:

IMDI Field	Value
Role	interviewee
Name	Peter
FullName	Peter Wittenburg
Code	W
FamilySocialRole	Unspecified

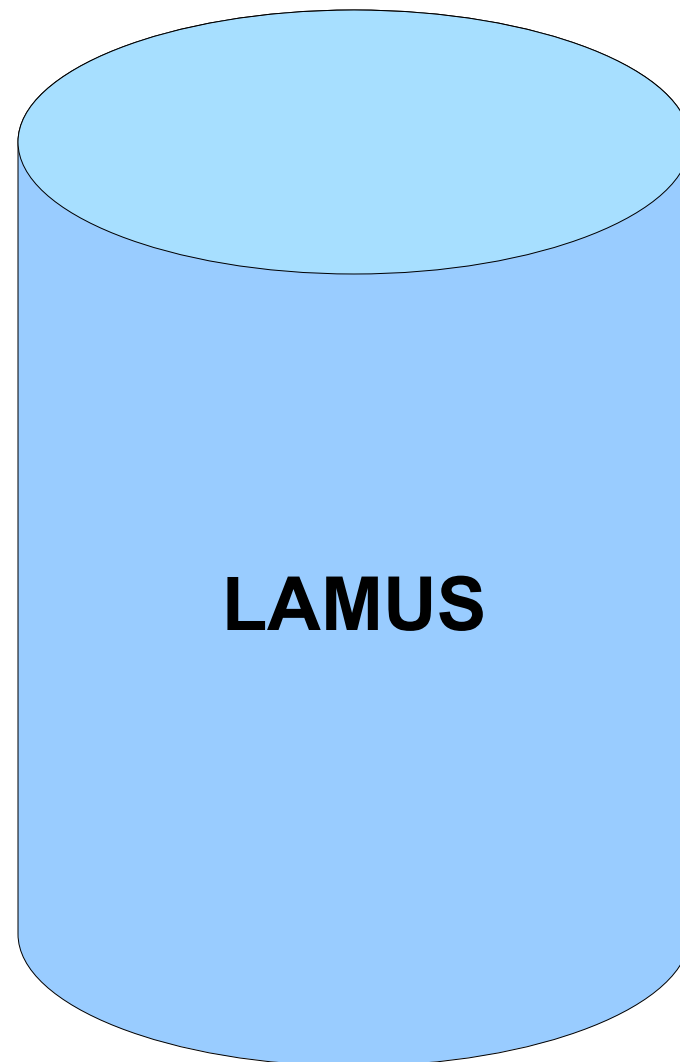
Below this table is a 'Selection' window, which is currently empty. Below that is another 'Selection(1)' window, which contains the following data:

IMDI Field	Value
Role	interviewee
Name	Peter
FullName	Peter Wittenburg
Code	W
FamilySocialRole	Unspecified
Languages.Description	
EthnicGroup	
Age	Unknown
BirthDate	Unspecified
Sex	Unknown
Education	university
Anonymized	true
Contact	
Description	



The Three Pillars of Archive Management

The Language Archive Management and Upload System

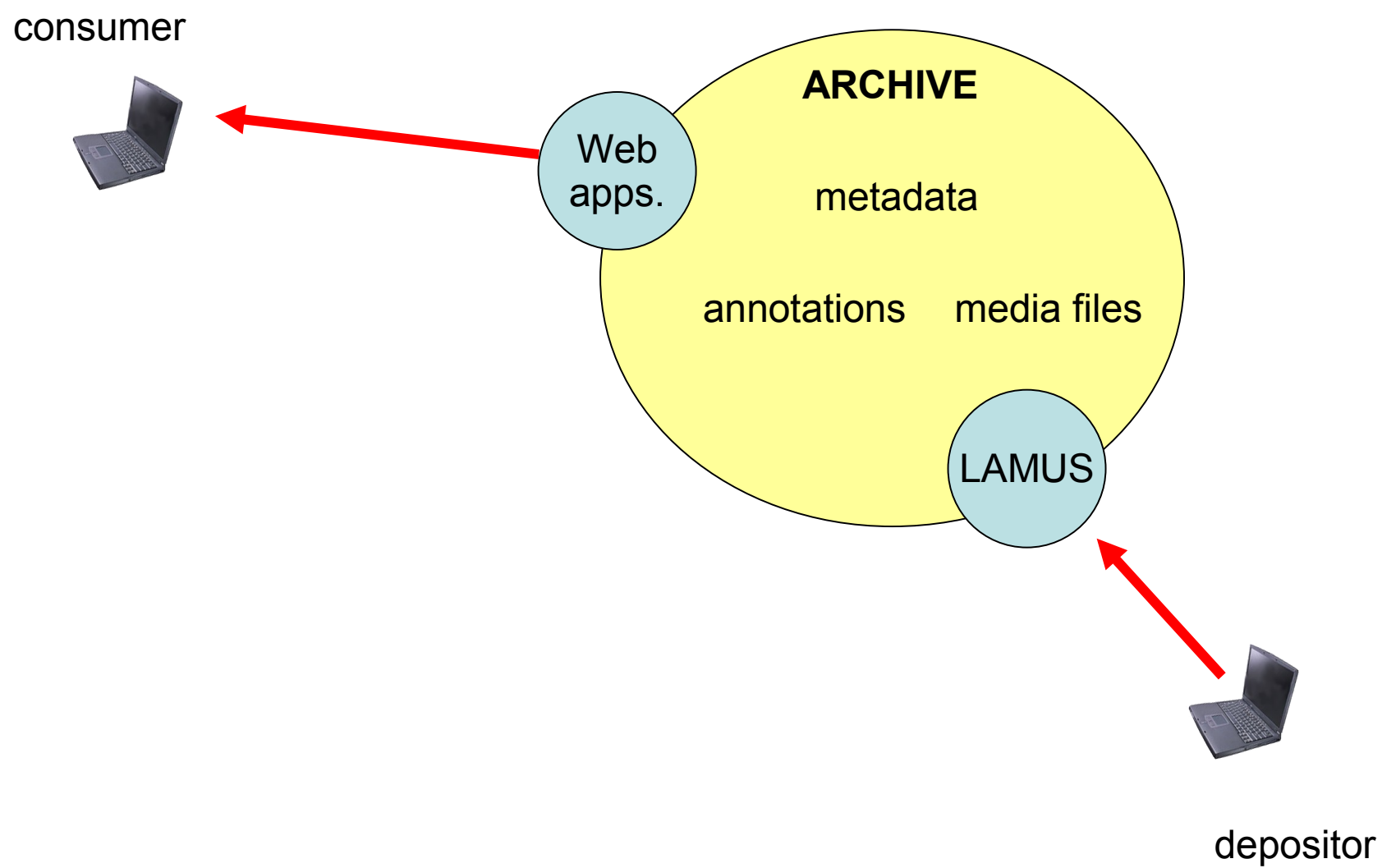




- Web-based operation
- Workspace principle
- Create / modify the (sub) corpus structure
- Upload / replace metadata & resources
- Gatekeeper function: not all data formats are accepted to keep archive consistent and future proof
- Assigning of URIDs
- Versioning system in the background



The Two Sides of Archive Access





The LAMUS Interface

Upload Resources

With the help of this page you can upload resources from your PC to the archive. Note that the uploaded resources will only be available to session-nodes that are linked directly from this node (or the parent node - as this is a session-node).

If you upload imdi session files that have external entity definitions in a file "imdi-sessions.imdi" you must upload this file also!

You have already used 0 MB. There is 10240 MB available.

Use the "Browse" button in the graphical window below to select the resources going to be uploaded. This can be done repeatedly to select multiple files. You can also select whole directories. After you see all the required files in the window click the **"Upload"** Button of the graphical window to start the upload process.

Name	Size	Directory	Modified	Readable?
------	------	-----------	----------	-----------

Upload 0% STOP

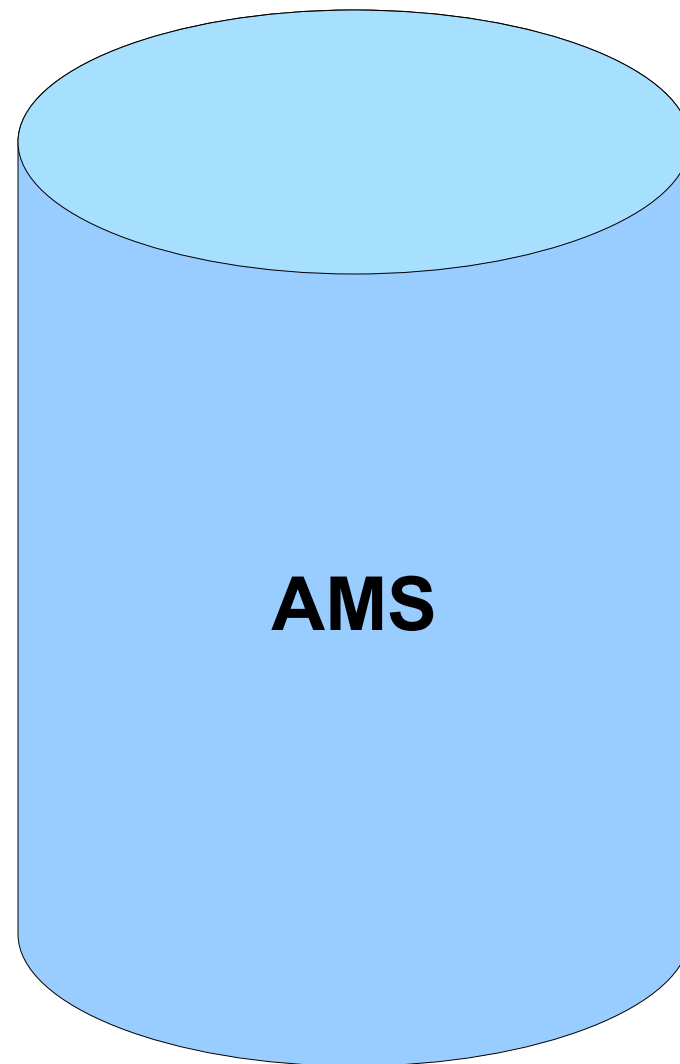
Buttons: Upload Files, Request Storage, Unlinked Files, Submit Workspace, Save and Logout, Delete Workspace, Help, Report a Bug, About

Applet lams.applets.wjhc.jupload.JUploadApplet started



The Three Pillars of Archive Management

The Access Management System





- **Metadata is accessible for everyone(OAI)**
- Resources are not always open for everyone
 - Sensitive information on actors
 - Ongoing research work with publication aims
- Access should be defined by researcher
- Different access rights for different users/user groups (access to video, audio and annotations can be set independently)



The AMS Interface

IMDI Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://corpus1.mpi.nl/ds/imdi_browser/

IMDI-Browser about user: alekoe@mpi.nl login logout

IMDI-corpora

- WelcomeToIMDIDomain.html
- index.html
- ALLA
- ANDES
- Bavarian Archive for Speech Signals (BAS)
- CORP-ORAL
- Coralrom
- DBD
- DoBeS archive
- ECHO
- ESF corpus
- Endangered Languages
- GTRP Corpus
- IFA corpus
- ILSP INTERA contribution
- Lablita
- Leiden Archives
- Lund Corpora
- MPI CGN
- MPI corpora
 - HTMLcorpusContent.html
 - Acquisition
 - Comprehension
 - Demo
 - Ams Demo
 - CLARIN
 - DOBES training
 - DoBeS Demo
 - PeWi corpus
 - kleve-route
 - Pim Levelt leaving MPI
 - Talks
 - YeleAnnexDemo
 - woonic_test_zone
- Language and Cognition

- view access privileges to LAT Resources 'PeWi corpus:' -

Focus overview on selected User/Group:

Rules of Node 'PeWi corpus'

- /MPI0#/MPI1#/MPI76399#/MPI76400# -

There are no licenses assigned to Node 'PeWi corpus'

- ▶ [Manage Node Licenses](#)

There are no privileges directly assigned to Node 'PeWi corpus'

- ▶ [Add new Rules](#)
- ▶ [Force Export](#)

Rules of Node 'Demo'

- /MPI0#/MPI1#/MPI76399# -

There are no licenses assigned to Node 'Demo'

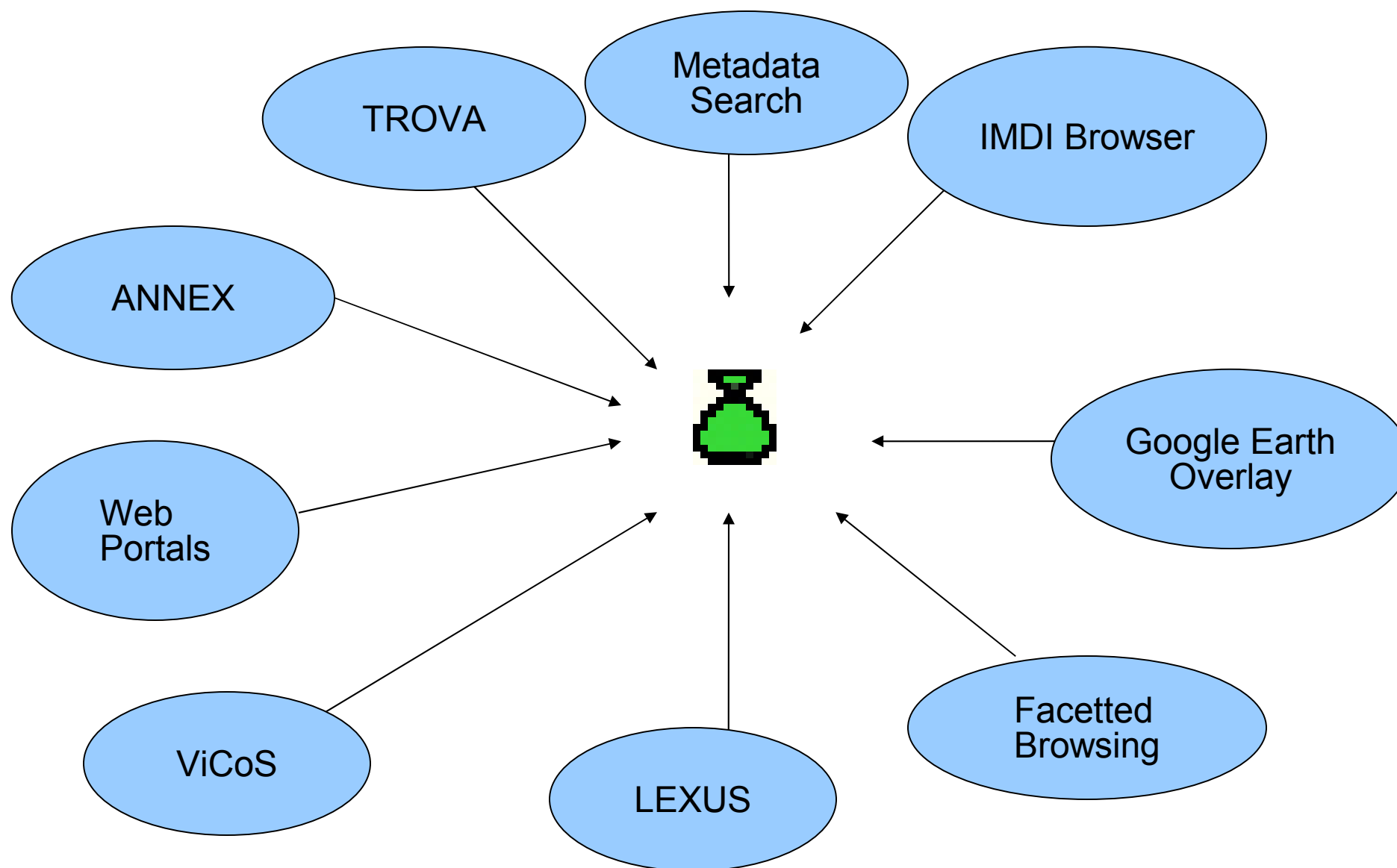
--- EVERYBODY ---

- Read Info Files Allow
- Read Annotations Allow
- Read Images Allow
- Read Audio Files Allow
- Read Video Files Allow

Done

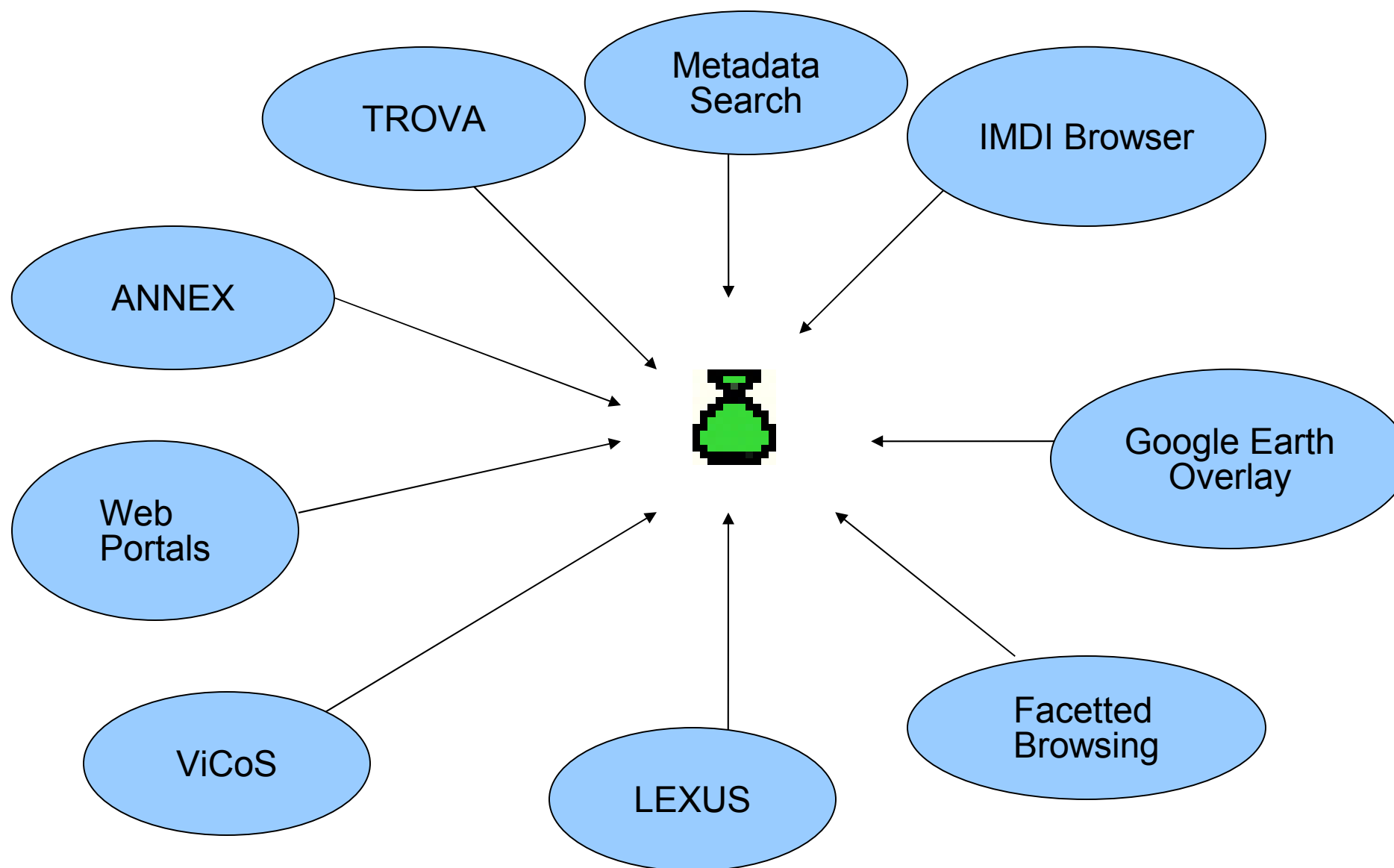


Accessing the Data





Accessing the Data



more on that in:

Ringersma, J., Zinn, C., & Koenig, A. (in press).

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

SDV - Sprache und Datenverarbeitung/International Journal for Language Data Processing.



Accessing the Data - ANNEX

- ANNEX can be called from IMDI Browser
- Video with synchronised annotations

The screenshot displays the ANNEX web interface. At the top, there is a navigation bar with 'ANNEX', 'manual', '?', and 'settings' on the left, and 'user: anonymous' and 'login' on the right. 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 conversation. Below the video are playback controls (play, stop, previous, next, full screen) and a 'Settings' button.
- Media information:** A panel displaying metadata: Resource: elan-example1.eaf, Media file: elan-example1.mp4, Elapsed time: 00:00:00:000, and Selected chunk information (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').
- Timeline:** A large panel showing a detailed timeline of the video. The timeline is divided into layers for different linguistic levels: K-Spch, W-Spch, W-Words, W-POS, W-IPA, W-RGU, and W-RGph. A tooltip is visible over the W-IPA layer, showing: Begin time: 00:00:00:780, End time: 00:00:04:090, and Text: səu ju: go aut əf ðə instɪtju:t to zə sɑnt ɑnə strɑ:t.

On the left side of the timeline, there are control buttons: 'Play selection', 'Clear selection', navigation arrows (|<, >|, <<, >>, <, >), and '+', '-' buttons. Below these are radio buttons for 'Play screen by screen' (selected) and 'Play continually', and a 'Tier text font' dropdown menu set to 'Arial Unicode MS'.



Accessing the Data - Metadata Search

- Metadata is always open and can be searched by everybody

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/ael1 reared
/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/rg3enamams1
/IMDI-corpora/Endangered Languages/DoBeS archive/Chontal/02	Society - Sociedad/Courting - Enamorando/rg5enamams2
/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



Accessing the Data – TROVA 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"/>	<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="*"/>



ELAN - **Time-aligned multimedia annotation**



- supports interchange with a wide range of formats (e.g. Shoebox/Toolbox, CHAT)
- uses XML, Unicode (easy to convert from/to, future-proof)
- supports a wide range of media formats (playback is delegated to Media Player, QuickTime or JMF)
- cross-platform (runs on Windows, Linux and MacOS)



- Annotations on multiple tiers
- tiers can be hierarchically interconnected
- annotations can be directly time-aligned to media or refer to another annotation layer
- can link in up to four different video files
- can also link in both a video & an audio file
- highly configurable
- complex search options
- in constant development, incorporating user feedback



ELAN Interface

The screenshot displays the ELAN software interface for the file 'elan-example3.eaf'. The 'Audio Recognizer' tab is active, showing the 'Silence Recognizer' settings. The 'Minimal Silence Duration' is set to 400 milliseconds, and the 'Minimal Non Silence Duration' is set to 300 milliseconds. A 'Start' button is visible at the bottom right of the settings panel. The main window shows a video of two men sitting and talking. Below the video is a timeline with a selection of 00:00:25.690 - 00:00:26.402. The audio waveform is visible, and the transcription tier 'K-Spch' is highlighted in red. The transcription text is as follows:

Tier	Text
K-Spch [7]	
W-Spch [15]	and then you go the this way eh ja to kleef
W-Words [97]	an th y go the this way eh ja to Kleef
W-POS [97]	co ad p v art de n part int pre n
W-IPA [15]	and ðen ju: go ðə ðis weɪ ə ja: tu kleɪf
W-RGU	



LEXUS - **Collaborative Lexicon Tool**



- web-based tool for creating & editing lexica
- workspace principle for multiple users
- 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



LEXUS Interface

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

Done

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)



ViCoS
-
Visualizing Conceptual Spaces



- 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



ViCoS Interface

Back Forward Reload Stop Home 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.../bug/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 ●

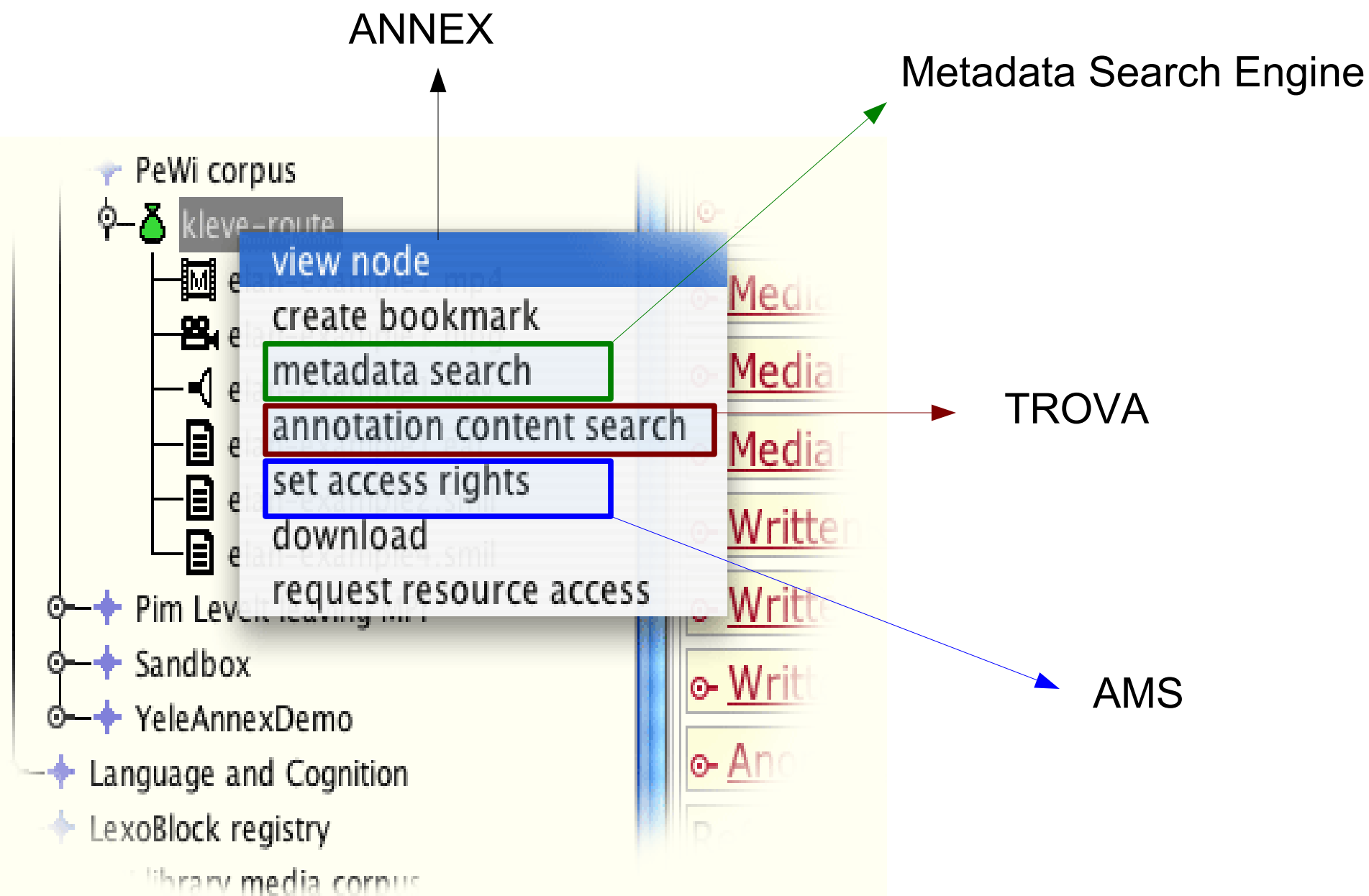
Modes:

browse	move
connect	delete
lexus	world
attach	detach
overview	refetch
save	colour

Relation Types:

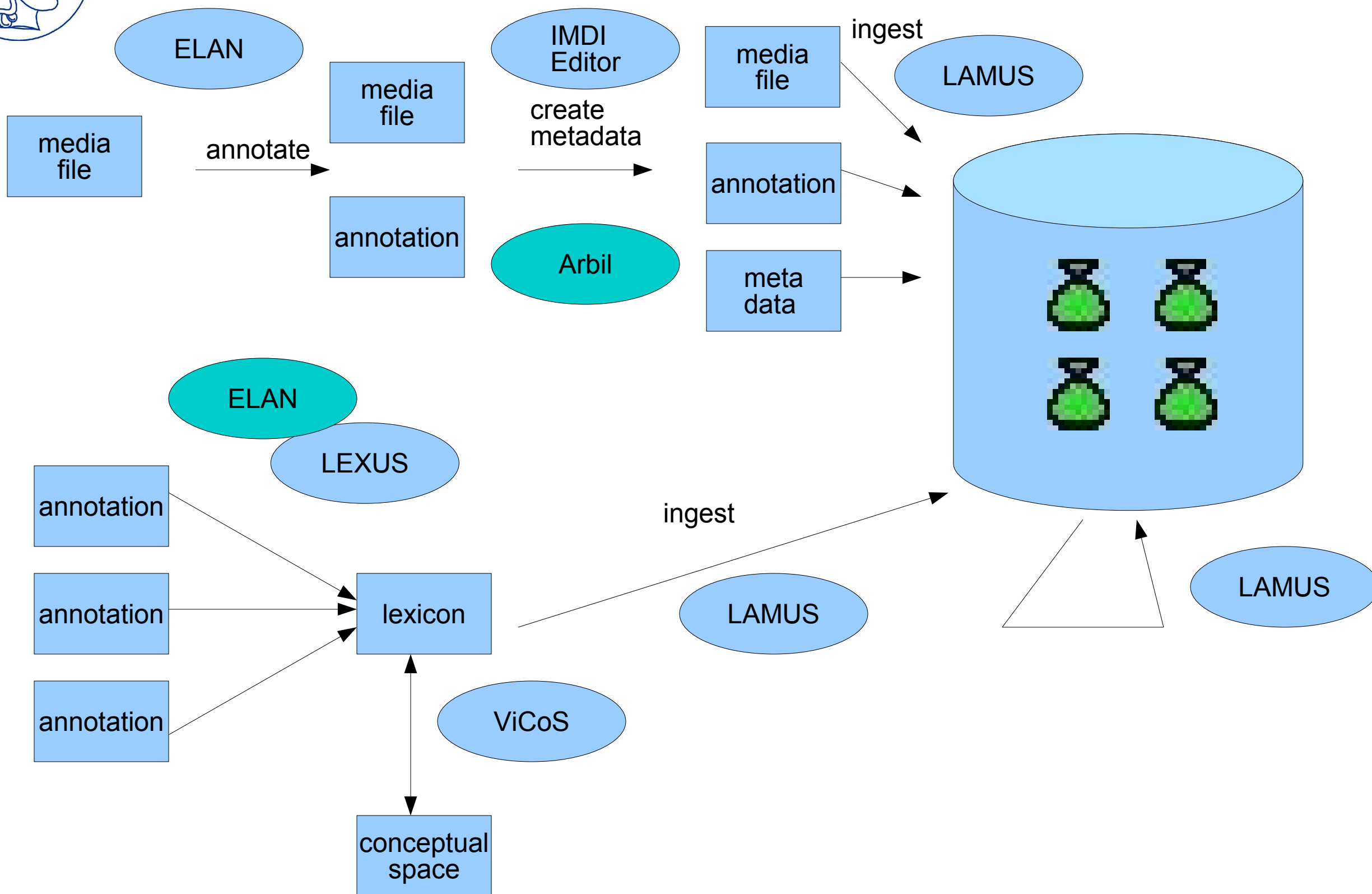


Interconnectedness – The IMDI Browser



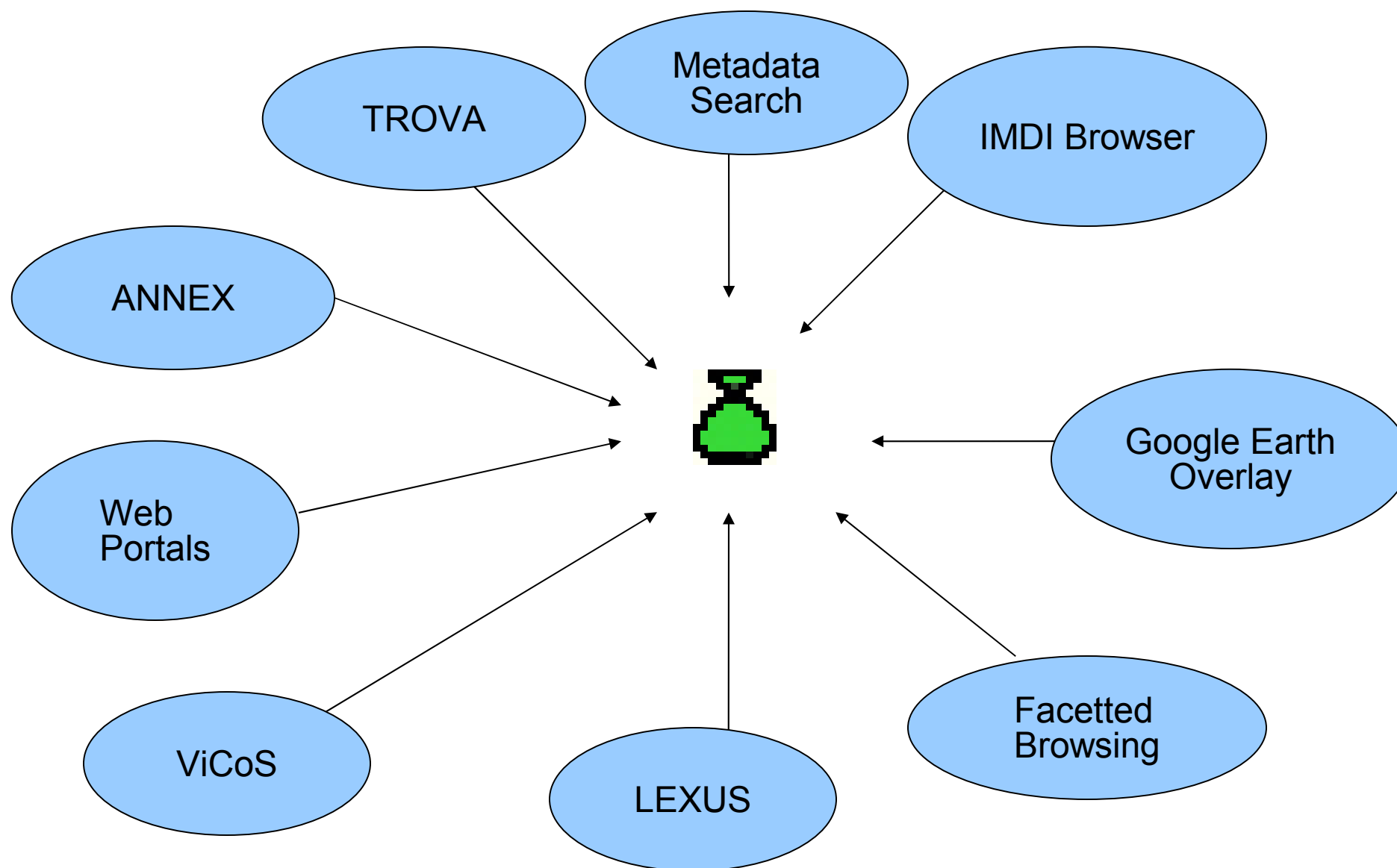


Interconnectedness - Getting Data into the Archive





Interconnectedness - Getting Data out of the Archive





Archiving instance, Max Planck Institute for Psycholinguistics

Archive managers: 3

Archive developers: 2

System manager: 1

Archiving software development: 4

Enrichment software development: 4





Archiving instance, Max Planck Institute for Psycholinguistics

Archive managers:

Alexander König

Jacqueline Ringersma

Paul Trilsbeek





Archiving instance, Max Planck Institute for Psycholinguistics

Developers:

Daan Broeder, Marc Kemps-Snijders,

Han Sloetjes, Claus Zinn,

Peter Withers, Mariano Gardellini,

Thomas Koller





Thanks!

Thank you for your attention!

