



Language Archives at MPI



Daan Broeder
Andreas Claus
Freddy Offenga
Romuald Skiba
Paul Trilsbeek
Peter Wittenburg

MPI for Psycholinguistics

DOBES Archive



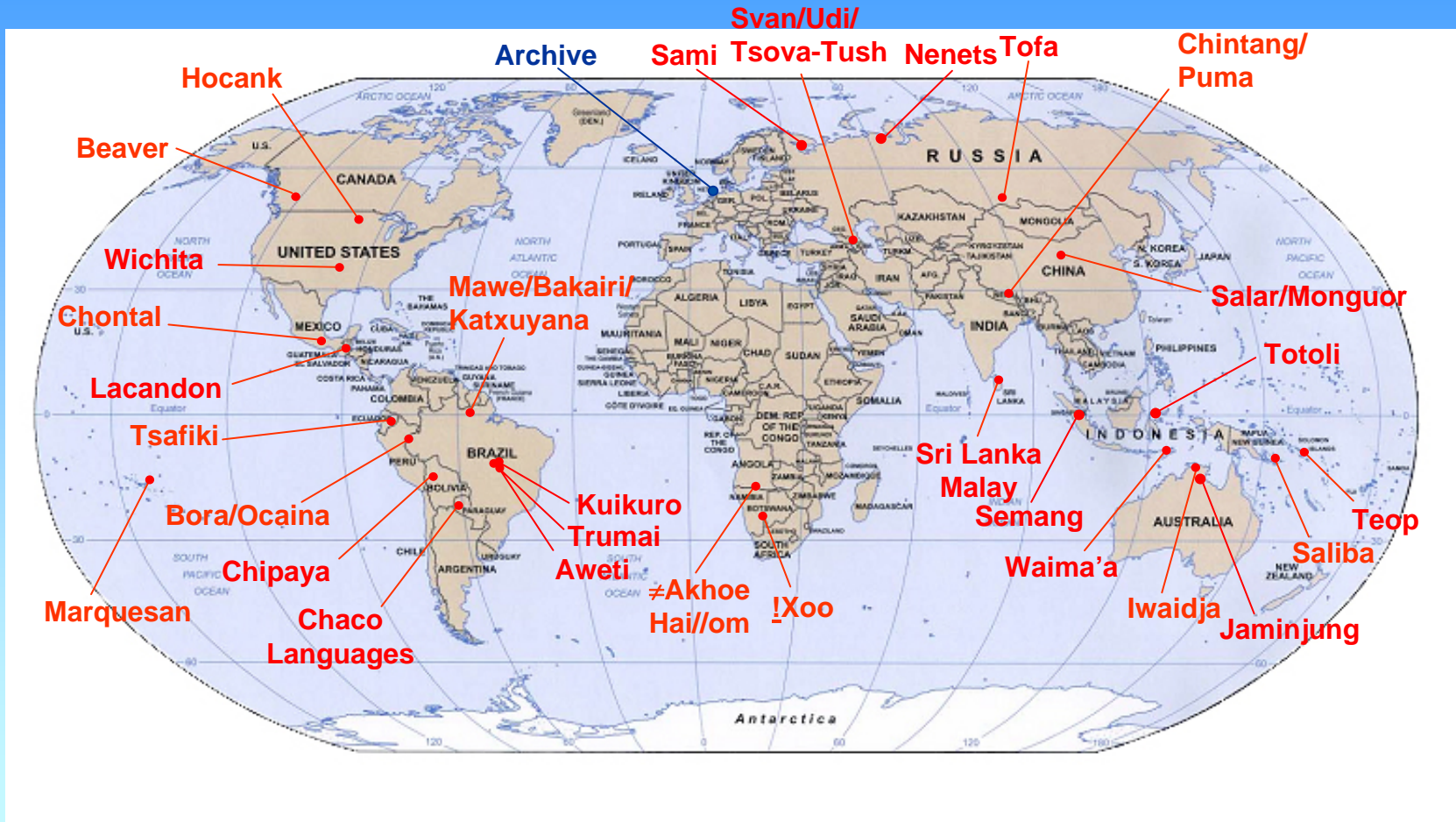
Topics



- what is in the archive – some examples
- why digital archives and what are their tasks
- principles of modern digital archives
- MPI archive architecture
- access management
- metadata organization
- resource upload
- archive federation
- services
- problems
- final statement



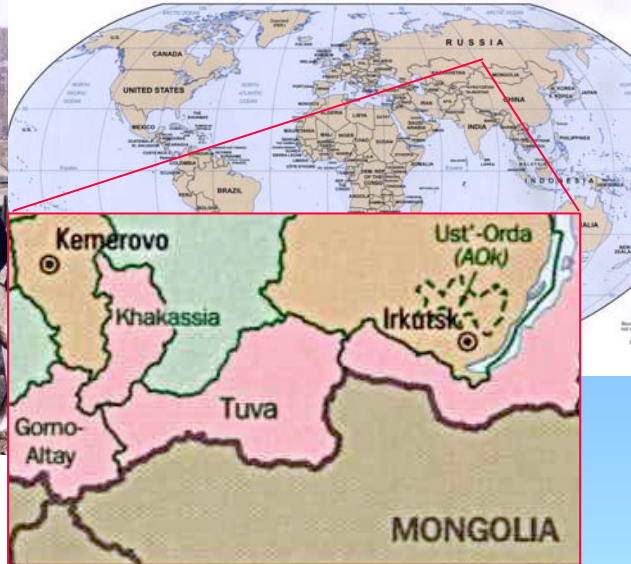
The DOBES Map



- 30 Documentation Teams (at MPI also 30 expeditions per year)
- 1 Archiving Team



Tofa, Tozhu, Tsengel Tuva, Tuha (Siberia)



- David Harrison (Yale)
- Brian Donahoe (Manchester)
- Sven Grawunder (Halle)

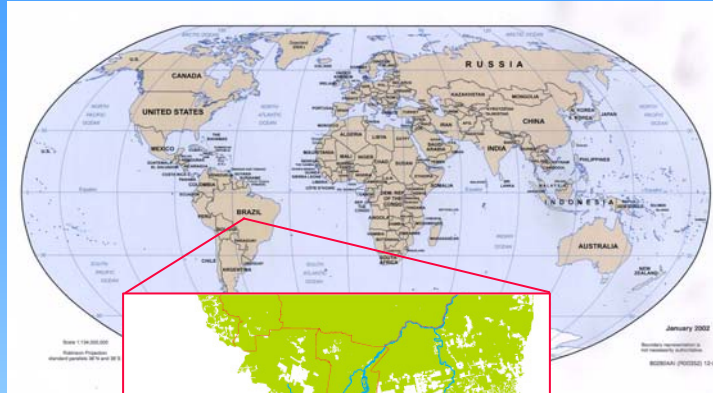
- Language—its structure and sounds.
- Oral folklore—texts, narratives and personal stories, belief systems, naming systems.
- Music—singing and sound mimesis.
- Traditional ecology—nomadism, pastoralism, hunting and reindeer herding



Shaman Ceremony



Trumai (Amazon)



Drawings



Video Clips
Annotated Media

Raquel Guirardello-Damian, Museu Paraense Emílio Goeldi

Steve Levinson, MPI

- *about 100 people*
- *about 51 speaking Trumai*



Yéî Dnye (Russell Island)



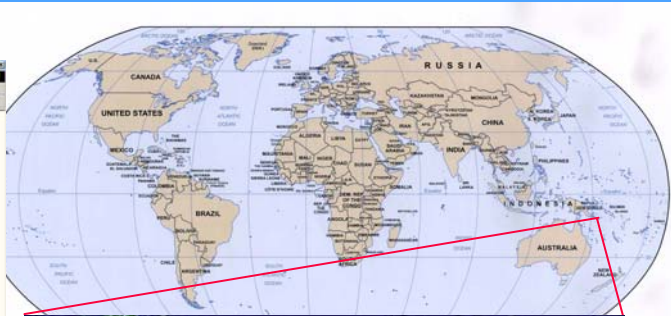
Browsable Corpus
hosted by the MPI for Psycholinguistics
(HTML version)

The Browsable Corpus gives you access to the IMDI-corpora housed at the MPI for Psycholinguistics. It allows you to browse through their hierarchical tree structure, whereby each node in this structure gives you access to further nodes. The nodes are characterized through metadata information, and they often have additional files associated with them: information files and, at the lowest level in the hierarchy, audio/video and annotation files. Metadata and information files are openly accessible, while audio/video and annotation files are usually protected for ethical and legal reasons. If you want to access protected files, please read the accompanying metadata files; they display the necessary contact information.

The Browsable Corpus is a preliminary test version for accessing XML-based IMDI-files with normal web-browsers. If you notice any errors, please contact the [CORPUS MANAGER](mailto:corpus@mpi.nl). Only browsers newer than Netscape 4.7 are supported.

If you don't see a tree on the left side of the page you may try our [download](#) of the IMDI browser. For displaying the tree on the left side you need to have Java (J2SE) installed ([see](#)).

Described Corpus



Lexicon Scheme Viewer
Sample lexicon for Rosset Demo 10/Feb/2006

Multimedia Lexicon

Lexeme: tpi:wee
Class: 1. 'sing sing' - see style of traditional sing, 2. verb, 3. person, singing
Example: 1. there's no singing, 2. tpi:wee daa:stokoo

Free Translation:
Note: 1. No clear relationship to here! This song style contrasts with you, it's used and other styles both usually and usually. Tpi:wee are only sung by men and boys, in performances that last all night. Performers wear special grass skirts and carry spears, 2. includes cockroach, stag beetle, wasps, ants, mosquitoes, white ants (and anything with ... wee) but includes flies, bees, case-worms, 3.

Video Clips



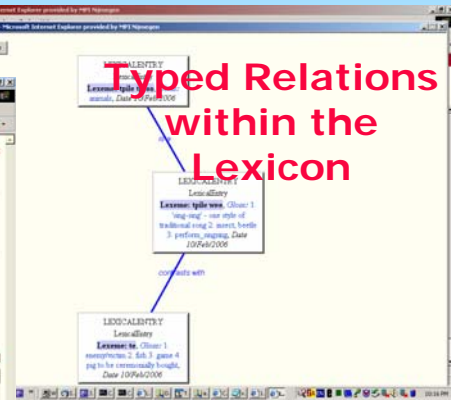
Photos

Annotated Media

File: r03_y00_v55.asf (MP3/240KHz)
View: [14] [hide controls] [CV]

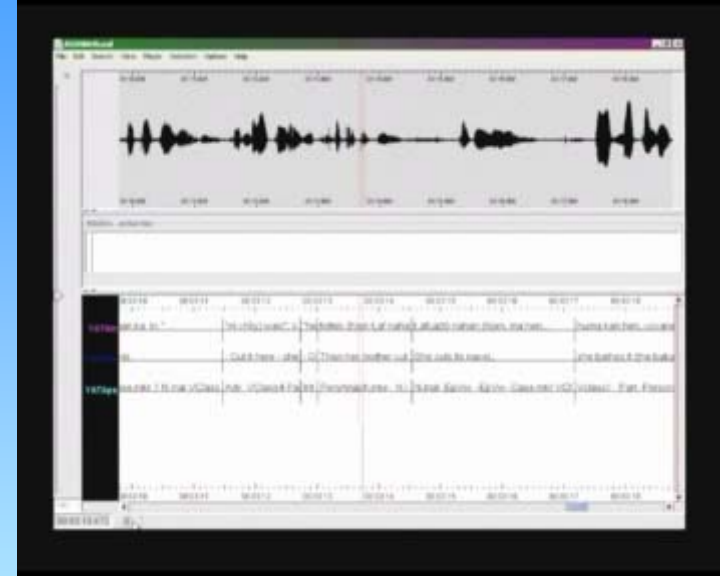
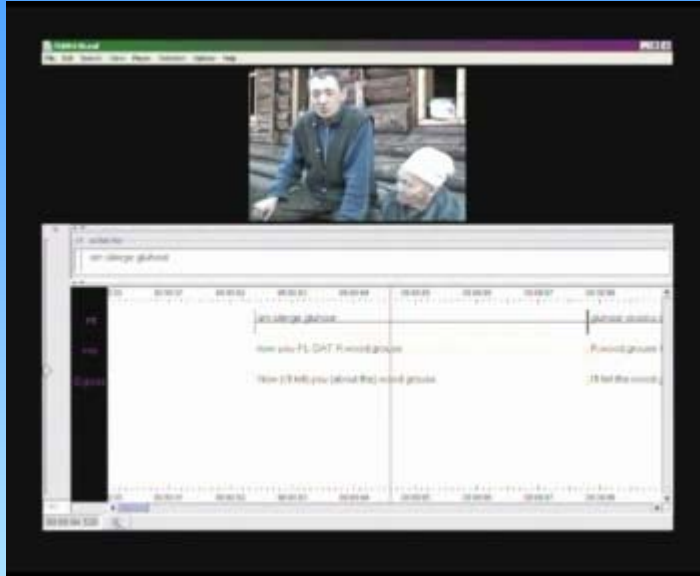
Class: [conts] [NONE]
hide times

just sing one! I also have never heard it (tricking?) **was singing** think it will go down here (in recorder) that's where you started at the **where were you até Yámbéyámbé** I was not staying close I was singing it beginning of the mbwaa. Ok, you started to sing VAK flower's opening it at K. I also made a mistake because of such a crowd sit down, respect seems to me just the same (as Myka tpi:we) sit down please I know this man I made a mistake, I went across because there were a lot of when you started the beginning of the mbwaa the second one. How people overtake me I never finished it we wanted to see you, you sing many have you in memory? this one is second You sing Myka's mbwaa that one sing the beginning part, start at the beginning I won't sing it. first This thing isn't just anything, it's going to record it's going to record people are coming closer. No, just sing it differently (not loudly) you go away, you Yámbéyá person OK, I haven't heard it, sing it in my hearing might sing it in church. You are 'trinking/tranging the Tpi:wee you tpi:wee that the beginning of the VAK? That's the beginning of the mbwaa yes? yes, the mbwaa It's the brain of the tpi:wee This one gives section (midnight) I started doing the Myka tpi:wee again because confirmation, then he will go. It's really blowing VAK! This guy is old here, I was seeing Yámbéyá as son of Kook Thasybra yes, I was eating going down (to foreground) I have present, he made record it's here no other foods no other boats have on this tpi:wee. What He went right around





What are annotated media files?





Why Digital Archives?



- several reasons
 - Dietrich Schüller: 80% of our recordings about cultures and languages are highly endangered!
inadequate Storage/Treatment (Media, Formats, PC, ...)
 - need data copying/distribution and continuous technology migration
requires discipline and formal procedures
 - change of technologies can become very expensive
- only central repositories can take care – Archives are such centers
- internationally accepted trend not only in our domain:
DOBES, AILLA, ELAR, PARADISEC, LACITO, ...





Modern Digital Archives



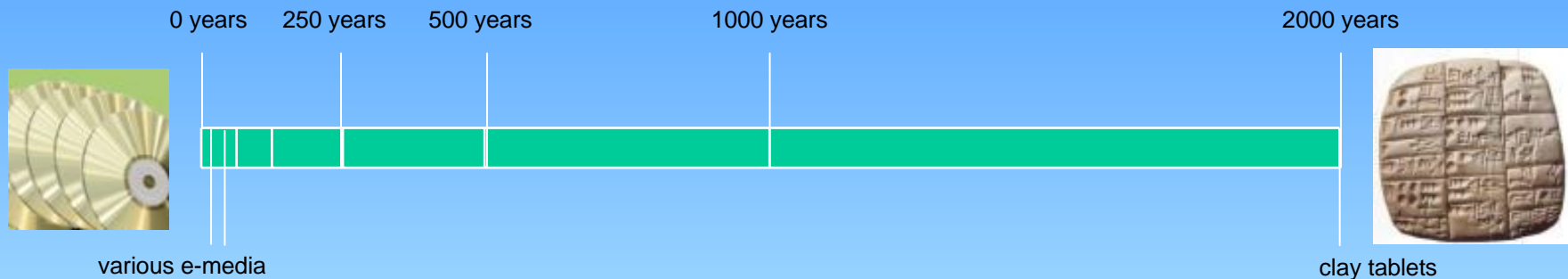
- traditional Archives
 - focus on long-term persistence of **physical objects**
 - access in general to be denied
- digital Archives
 - physical objects (tapes, CD-ROMs) are not relevant anymore
 - **content** has to be preserved
 - why this revolutionary change?
 - copies are lossless (be careful with compression)
 - copies can be created cost-effectively
- tasks of modern digital Archives
 - long-term preservation of the content (Migration, Distribution)
 - give access to the content
 - content enrichment without altering original resources
 - reliable access management



Long-term preservation: how?



- current storage media are not appropriate for long-term preservation



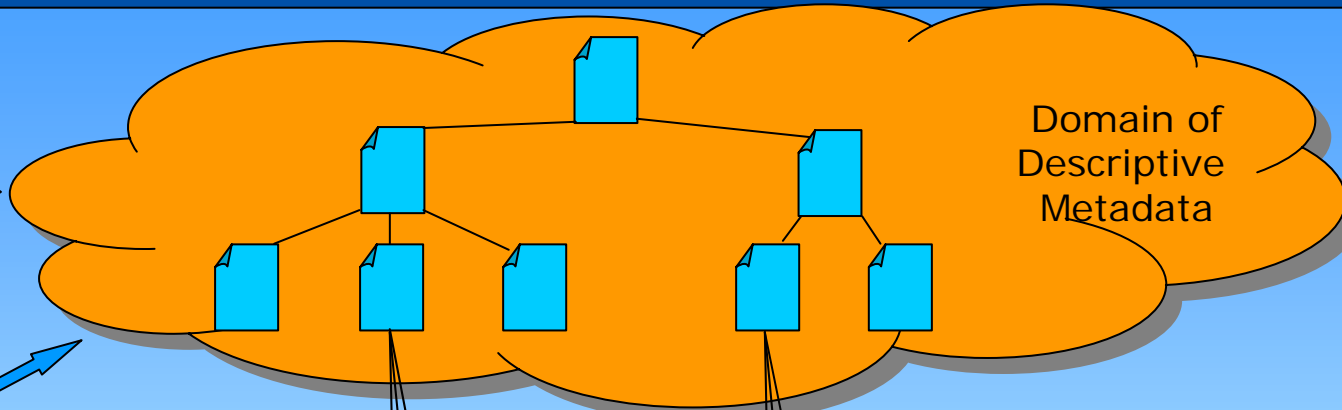
- increasing chance of data survival:
 - continuous migration (new technologies)
 - worldwide distribution of data
 - for DOBES -> 7 copies of the data in NL and DE, institutional backing from Max Planck Society for 50 years
 - within DAM-LR & DELAMAN -> further world-wide distribution
 - > take care of ethical principles, access restrictions must remain
 - take care of bit-stream survival – interpretation problem remains
 - **low costs for migration will be important (-> coherence)**



Principle I - Metadata



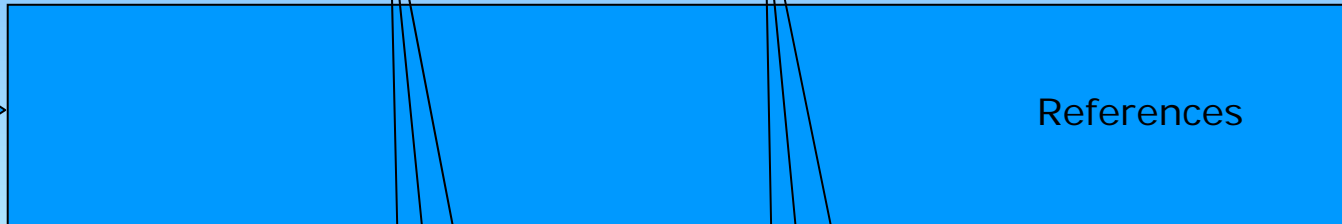
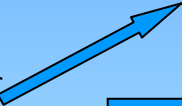
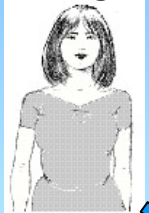
User



Domain of Descriptive Metadata

- open
- virtual
- linguistically ordered
- stable
- validated

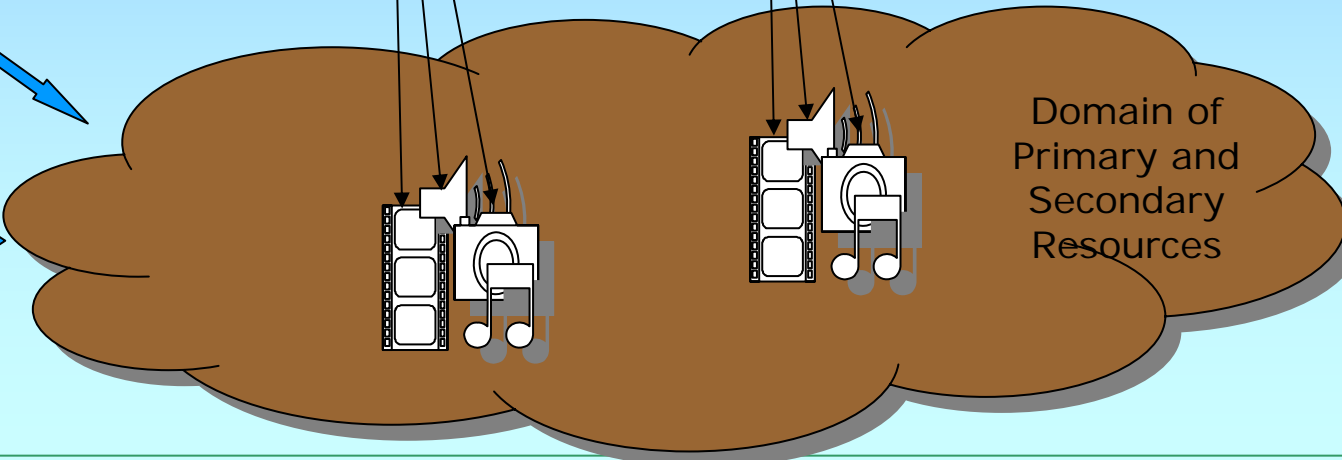
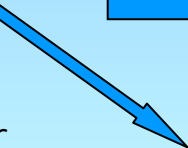
Corpus Manager



References

- consistency checks needed

System Manager



Domain of Primary and Secondary Resources

- restricted
- physical
- technically ordered
- subject of changes

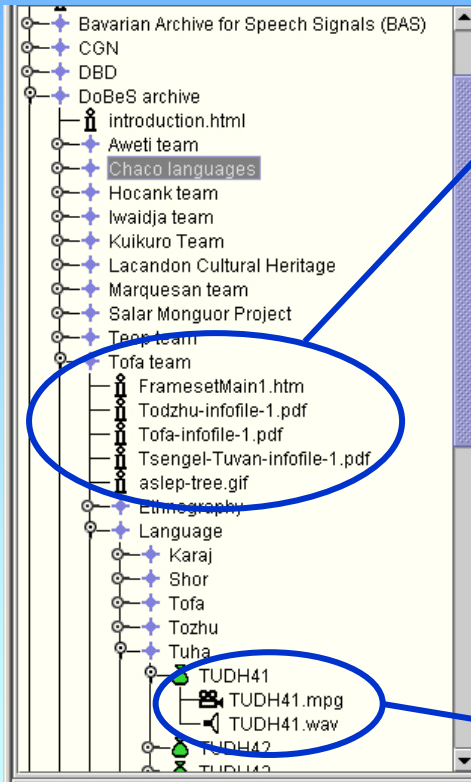


Principles II - Bundling

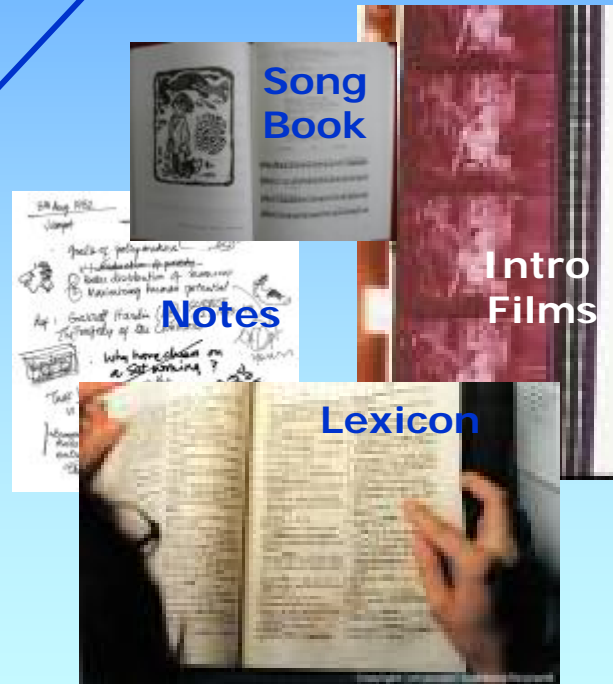


as many layers as researcher needs

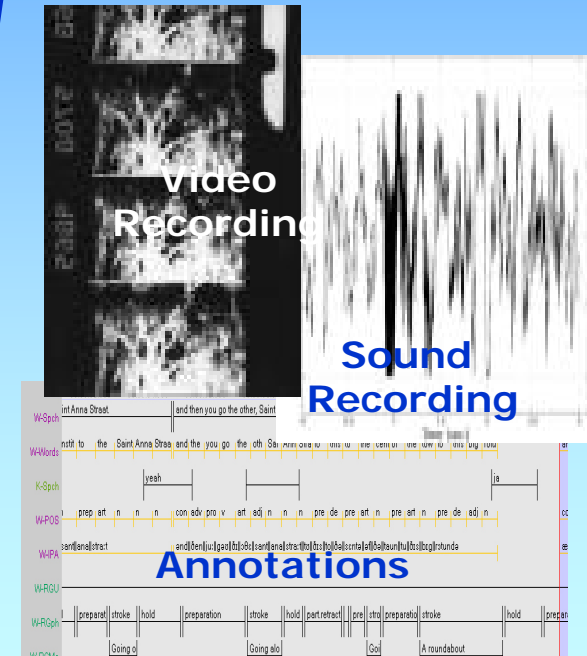
Archive Organization



Language Layer



Session Layer



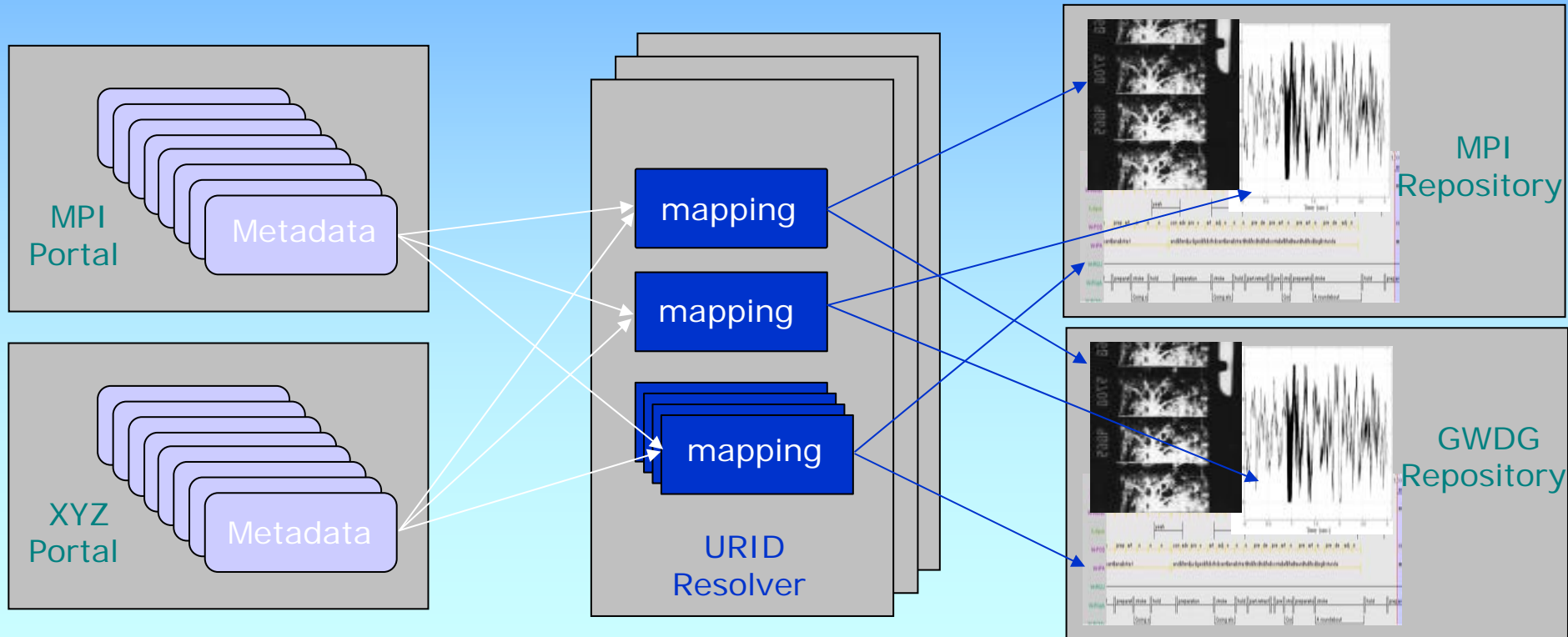
Bundling of information is crucial



Principles III – unique identifiers



- have to separate object and its instances
 - need Unique Resource IDs (comparable to ISBN number)
 - and a stable and reliable “Resolving” mechanism

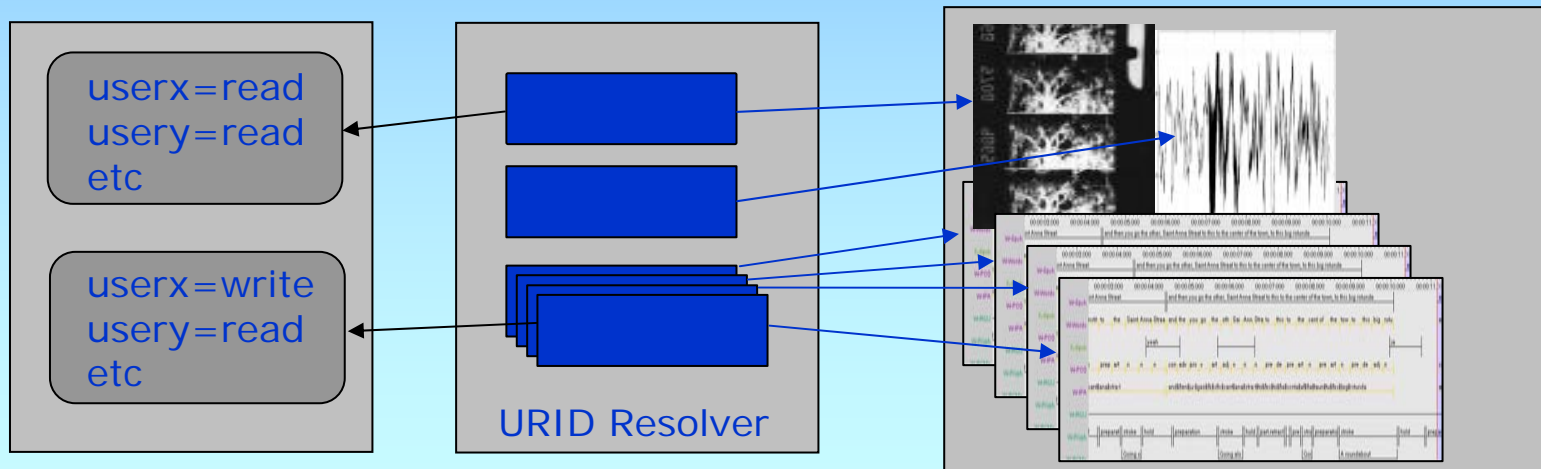




Principles IV - Versioning



- need Versioning
 - annotations etc will be changed – but nothing may be deleted!
 - science is a dynamic process
 - every version has a separate URID

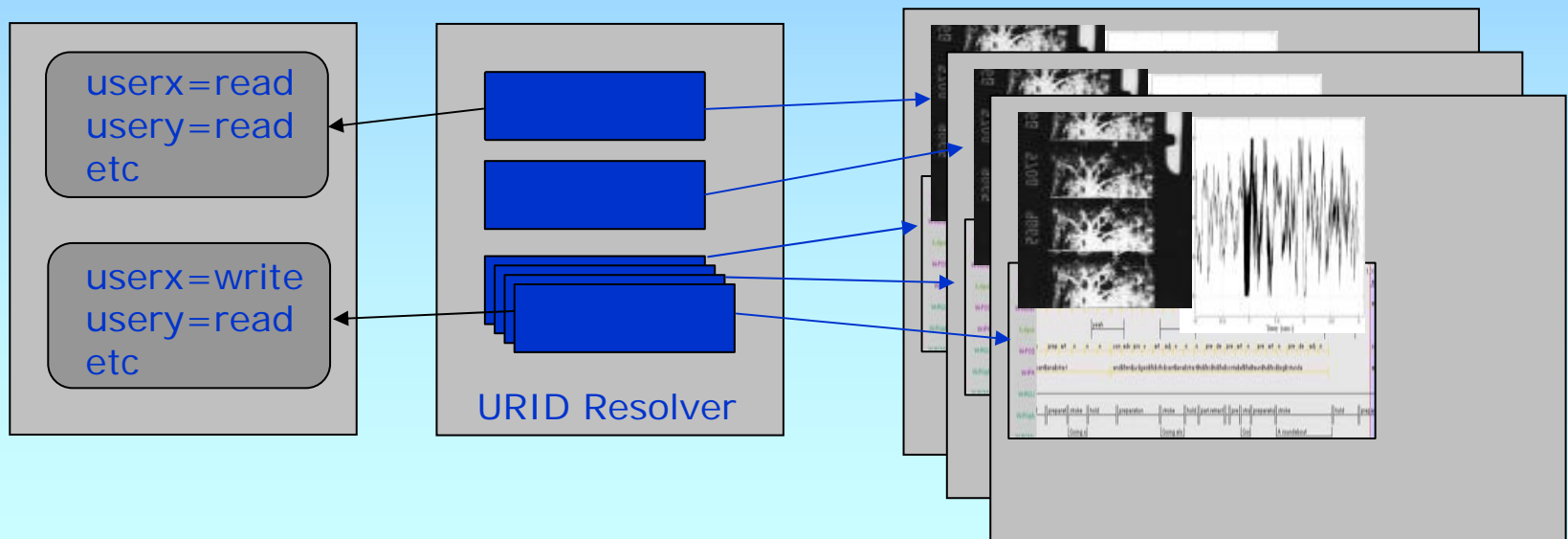




Principles V – Authentication&Authorization



- authentication and authorization to be separated
 - URIDs are central link to authorization information
 - authorization is coupled to objects
 - authentication with normal rules (password, encrypted)
 - careful handling of IPR of great importance
 - need to have space for policies, procedures, declarations etc





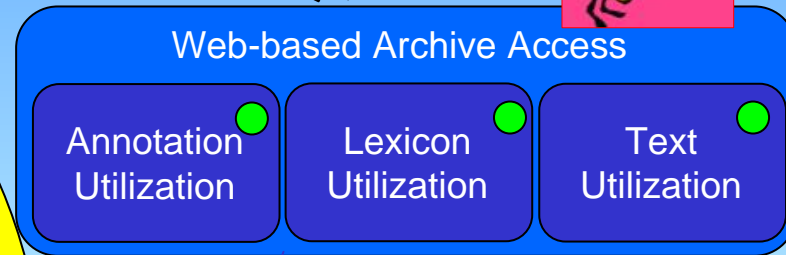
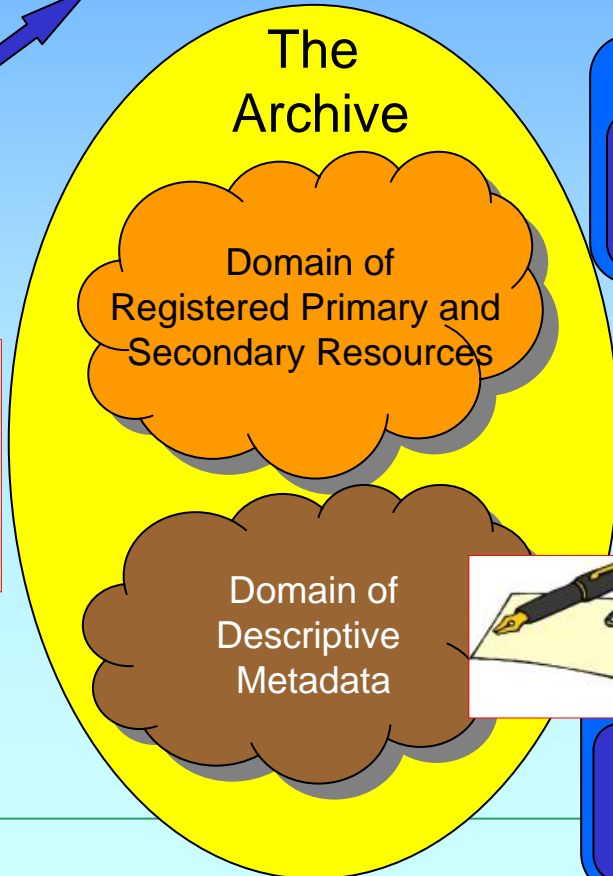
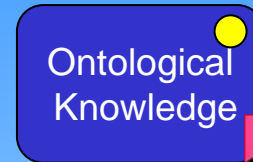
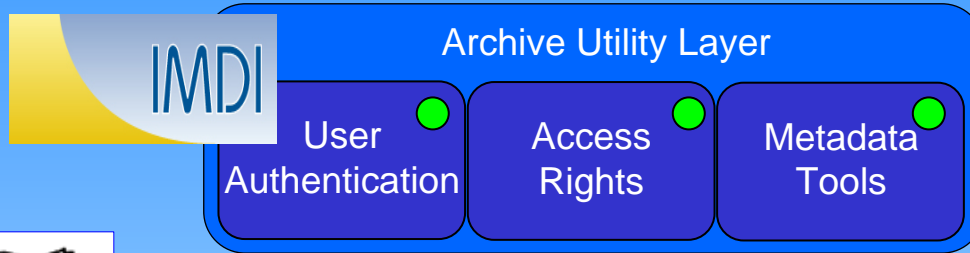
Principles VI – Formats



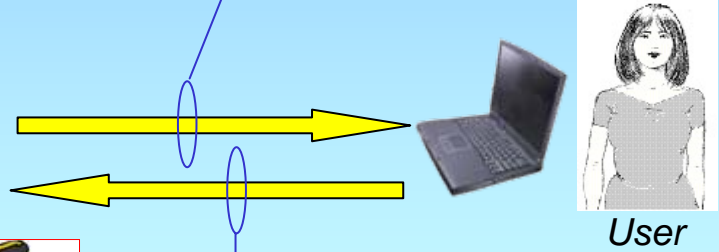
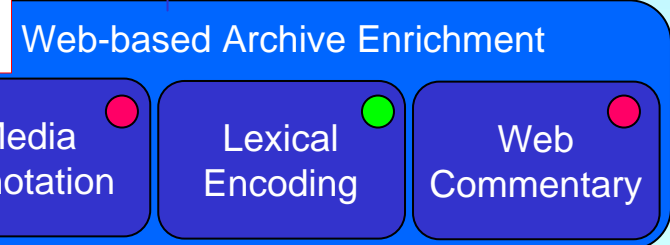
- only open, well-documented and widely used formats (encoding standards) should be used in the archive
- where possible generic schemas should be the basis
 - in DOBES strong recommendations for
 - JPEG/TIFF/PNG, MPEG2, Linear PCM, UNICODE, XML
 - Plain Text, HTML, PDF possible
 - at MPI less restrictive (therefore great danger with some types)
 - for presentation purposes also MPEG1/4, MP3, HTML
- archived objects have to be stored in a neutral way and accessible as individual objects including metadata that is important
- no encapsulation is permitted (databases, zip files, etc.)



MPI Archive – Architecture

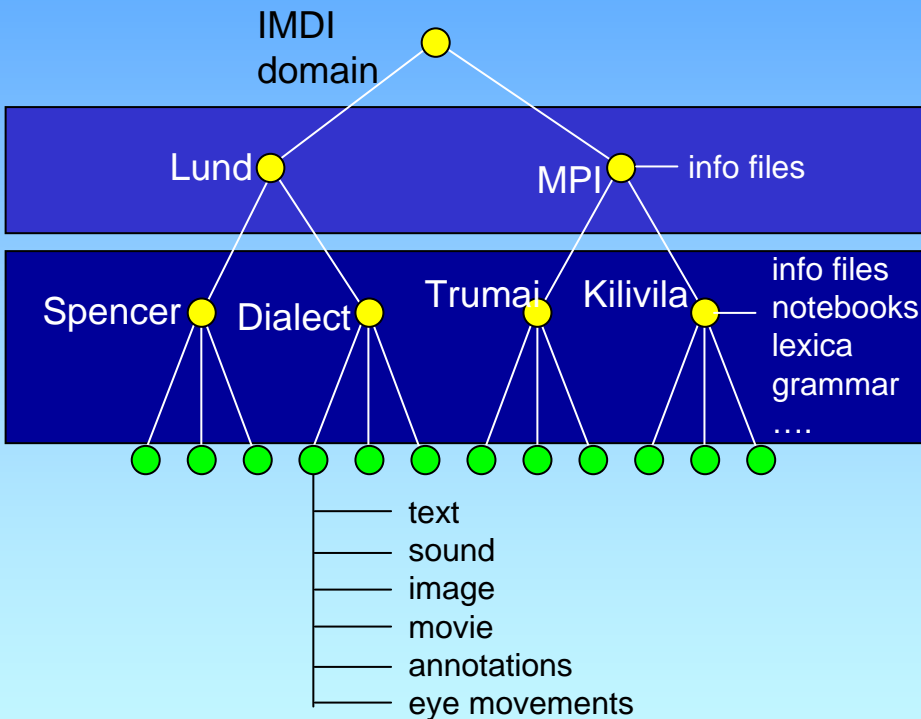


Primary Resources:
 Texts
 Images
 Sound
 Movies





Metadata Based Organization



- IMDI metadata is the glue that keeps related resources together
- every user can choose his optimal organizational form
- metadata is useful for management
 - by depositor
 - by archive manager
 - easy location of resources
 - easy check of state
 - easy sub-archive copying
 - etc
- IMDI is a linked domain of XML-files
 - can combine archive and local resources on notebooks
 - can combine different archives

IMDI

is mature – developed over 5 years



Basic Rights



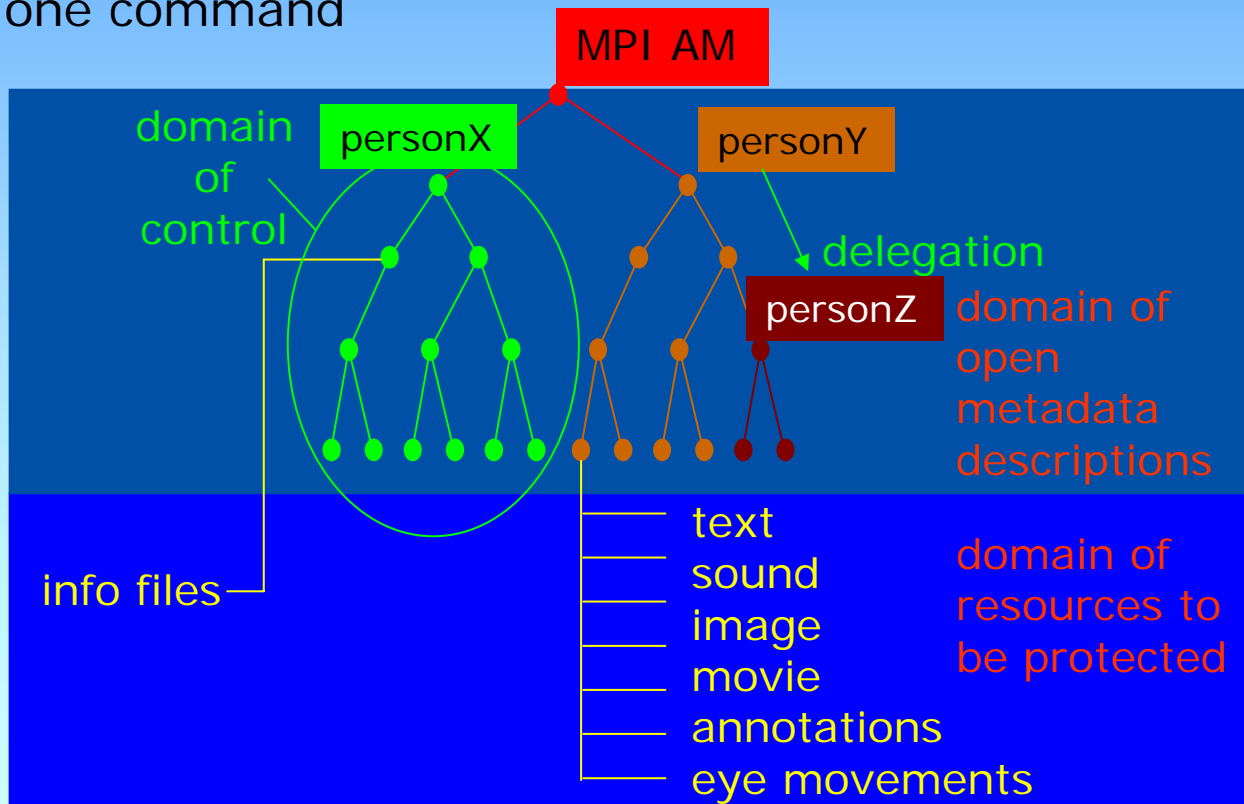
- only archive managers have all permissions (inc. delete)
- of course system managers have root passwords
- all bound to correct behavior (Code of Conduct, contracts)
- within MPI no operation on corpus may be carried out without archive managers permission
- also directors cannot overrule the signed agreements
- archive claims the right to archive
- copyright remains with the creators (researchers, consultants)
- depositors have the right to access their data
- can define rights for language community members
- for MPI the depositor is the key in all respects



Access Management System



- AMS is based on IMDI
- per sub-archive policies can be defined (DOBES: CoC and usage declaration acceptance)
- select an archive node and define rights for resource types with one command





Resource Upload Problem



- two options:
- manual integration
 - exceptions are easy
 - too many teams for archive management to handle (~60)
- software controlled integration
 - exceptions are difficult
 - users can do it themselves (?)
- a mixture will most likely become the reality



Resource Upload with LAMUS



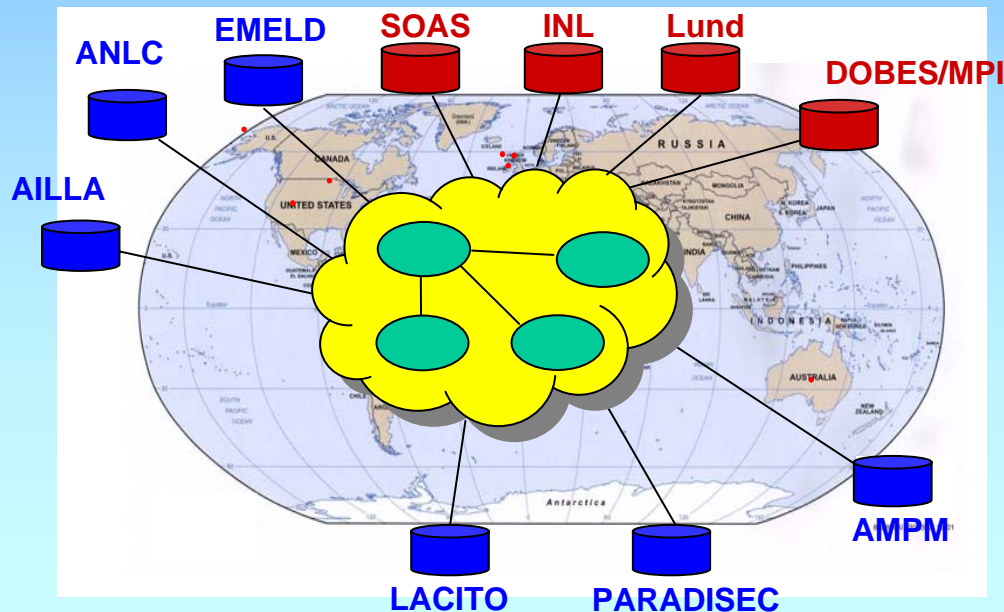
- LAMUS features
 - web-based operation
 - request of a workspace that is normally given for a period of time
 - specification of an accepted upload node (archive anchor)
 - extend and manipulate the corpus structure
 - upload metadata descriptions
 - upload any type of resources (**configurable format control**)
 - create a linked sub-archive in the workspace and integrate this into the archive
 - before integration a number of checks are carried out
 - LAMUS is handling a list of accepted file types, parsers can be used to validate, not accepted types and invalid files can be rejected
 - per default the integrated resources are not accessible except for the depositor
 - it generates content indexes when new resources are integrated to enable fast searching by utilization applications



Archive Federations



- we are moving towards a distributed A&A scenario
 - DAM-LR as a GRID test bed combining U Lund, INL, SOAS London
 - single sign-on, single identity principle
 - seamless navigation in joined archives as goal
 - DELAMAN network as perspective
 - probably turn over to LDAP and Shibboleth





Service 1



you can deposit your data at the MPI

- there will be an agreement with the depositor
- the depositor will have the right to access his/her data
- the depositor will have the right to control access to the data
- their has to be a positive attitude towards research collaboration

- 4 projects already make use of this service
 - Guanchet data
 - Narrangansett data
 - Dutch Bilingual Database project
 - ECHO Sign Language Corpus

- MPI offers fellowships to prepare and organize the data in case of important data
- MPI wants to help to preserve endangered material
- in case of interest call Paul Trilsbeek or Peter Wittenburg



Service 2



we can install the archive machinery at your site

- install a full-fledged LAMUS system and utilization tools
- did so recently in Lund – including all adaptations archive was online within 3 hours
- have agreements with 3 other sites in 2006
- have to give accompanying training courses
- goal to have a simple installer hopefully to come in 2006

- you are the full master of the archive system
- a service account for MPI guys makes sense for fast problem solving

- your data is immediately part of the IMDI domain



Problems



- coherence is important for management and utilization
 - however, researchers come with a wide variety of formats
 - they are often not aware of the relevance of standards
 - they argue from tools (which is understandable)
- often the workflow process is not smooth
 - problems with automatic tape handling (photos in video stream)
 - wrong digitization/capturing formats or parameters
 - cutting information is not correct
- researchers don't take the time to create metadata
 - 2/3 of all data of MPI are in a pre-processing stage
 - i.e. they are not tagged even with simple metadata



Final Statement



- Many language resources “highly endangered”
- Archiving them in specialized digital archives gives a higher chance of long-term preservation
- Good metadata organization is essential for archiving and utilization
- A limited amount of well-documented, open file formats and encodings in the archive makes migration to new formats feasible in the future
- An “archive content management system” can reduce work load for archive managers and will allow an archive to accept more data
- Distributed archives will further enhance the chance of long-term data preservation and will reduce dependency on individual institutions



LAMUS Screenshots



LAMUS - Language Archive Management and Upload System

http://corpus1.mpi.nl/jkcc/lamus/lams.jsp

Beaver Archive
Data
General introductory comments
Studies

Language Archive Management and Upload System

How to use LAMUS?

The LAMUS screen is divided into three sections. The text you are now reading is in the main section. Here you will find all the information to interact with the system.

In the left section you will see the current corpus tree (workspace) you are currently working on. To select a node from the tree you click on it with the left mouse button. Once a node is selected click the right mouse button to get the tree menu. With the help of this menu you can view, add, modify and link nodes.

At the bottom section are the buttons to select the different LAMUS functions. These functions are described below:

Upload Files

Transfers language resources and IMDI files from your local machine to the archive.

Request Storage

To make it possible for you to upload files to the archive corpus management needs to reserve space for it. With this function you can request the space you require to upload your files.

Unlinked Files

Shows the nodes and resources which are not linked into your corpus tree. Uploaded files and unlinked tree nodes are listed here.

Submit Workspace

When the corpus manipulation is finished and the resources are uploaded and linked into the tree you need to submit this request to ingest the new corpus into the archive. As long as the ingest request is not submitted other users will not be able to see your changes.

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

Applet treeviewer started



LAMUS Screenshots



The screenshot shows a web browser window titled "LAMUS - Language Archive Management and Upload System". The address bar shows the URL "http://corpus1.mpi.nl/jkc/lamus/lams.jsp". The page content is as follows:

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 1024046 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.

Browse... Remove Selected Remove All

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

Upload 0% STOP

When you do not see a graphical window above this text or the application does not work on your computer please use the file browsers below this text. First enter the amount of files you want to upload and click "Change". The chosen amount of upload boxes will appear. Choose the files by clicking the "Browse..." Buttons below, then click the "Upload" button.

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



LAMUS Screenshots



The screenshot shows the LAMUS web interface. The browser address bar displays <http://corpus1.mpi.nl/jkc/lamus/lams.jsp>. The page title is "LAMUS - Language Archive Management and Upload System".

On the left, a tree view shows the hierarchy: Beaver Archive > Data > General introductory comments > Studies > culture. A context menu is open over the 'culture' node, listing actions: view node, add corpus node, modify node, replace node, delete node, unlink node, link node, link info file, link corpus node, link session node, and link external node.

The main content area is titled "Upload Resources". It contains the following text:

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 1024046 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.

Below the text is a graphical window for file selection. It has a "Browse..." button, "Remove Selected", and "Remove All" buttons. Below these is a table with columns: Name, Size, Directory, Modified, and Readable?. The table is currently empty.

At the bottom of the graphical window is an "Upload" progress bar showing 0% and a "STOP" button. Below the progress bar is a text area containing "Action : Browse...".

At the bottom of the page, there is a row of buttons: Upload Files, Request Storage, Unlinked Files, Submit Workspace, Save and Logout, Delete Workspace, Help, Report a Bug, and About.