

DOBES/MPI Archive - architecture -

Paul Trilsbeek, Roman Skiba, Peter Wittenburg
MPI for Psycholinguistics



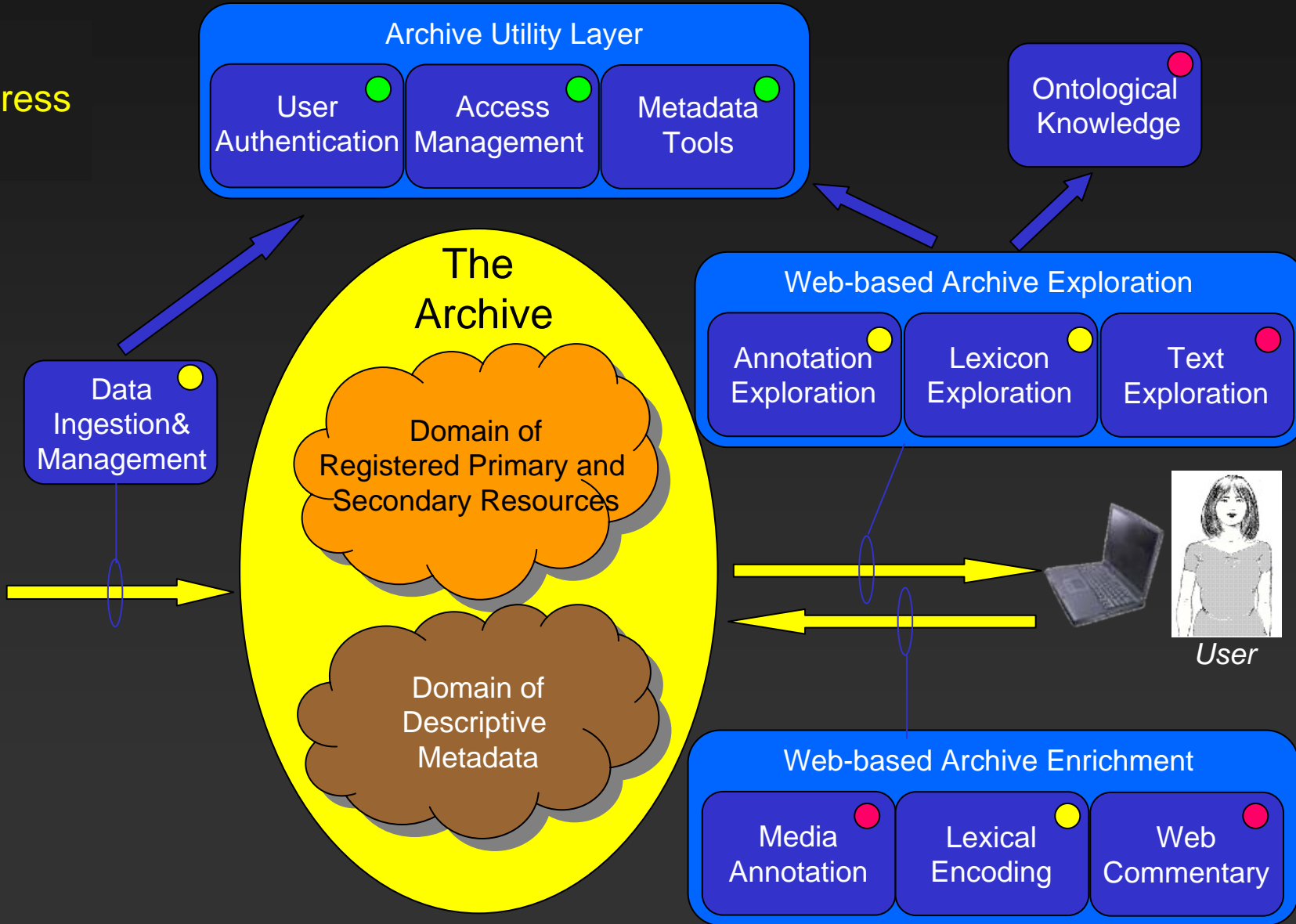
- we take almost all types of input if it is part of agreements (including even structured WORD)
- however, there is a strong recommendation for a limited number of formats – otherwise the job is not tractable
- only for metadata we have a strict policy
- DOBES follows the “program approach”

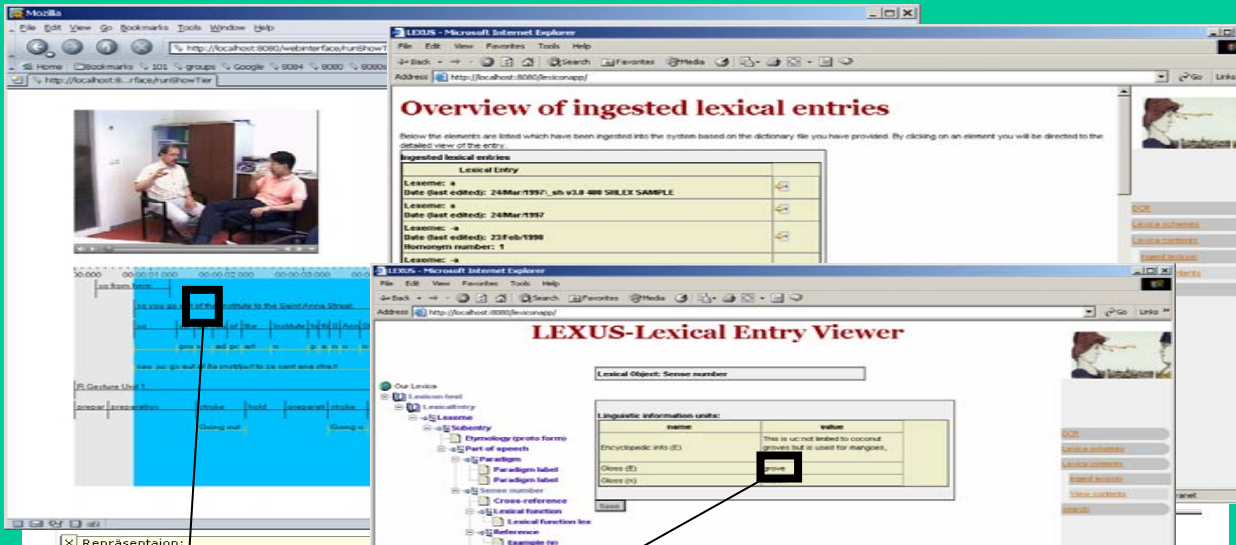
level of unification	XML	Schema	Elements Attributes	Value Range
project approach	X	X	X	X
programme approach	X	X	-	-
store all approach	-	-	-	-



- in the archive we only support a very limited number of encoding standards and formats
(UNICODE, lin PCM-wav, JPEG/PNG/TIFF, MPEG2/1/4, XML, HTML, plain text)
- structured data should be schema based
 - EAF annotation schema which turned out to be powerful enough
 - LMF lexicon schema which will become the new ISO standard
 - IMDI metadata schema which stabilized during the years
- MPEG2 as video archiving format although not the final solution
- various presentation formats
MP3, MPEG1/4, SMIL (subtitled video) etc

- done
- in progress
- to start

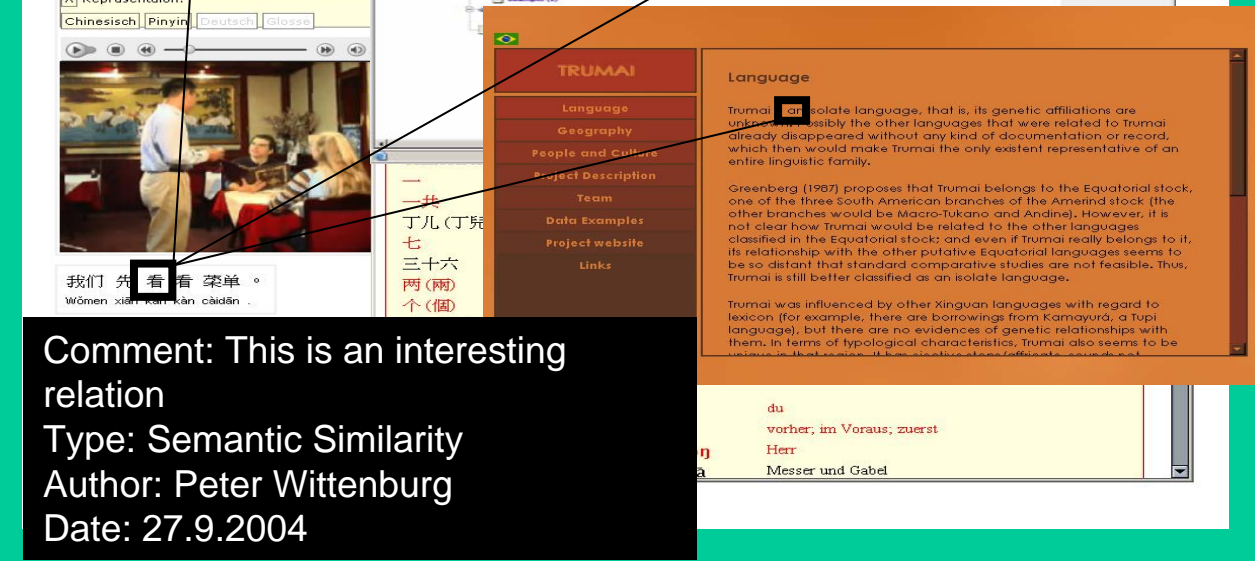




The screenshot shows a web browser with two windows. The top window is titled 'Overview of ingested lexical entries' and displays a table of lexical entries. The bottom window is titled 'LEXUS-Lexical Entry Viewer' and shows a detailed view of a lexical entry for 'Seesaw number'.

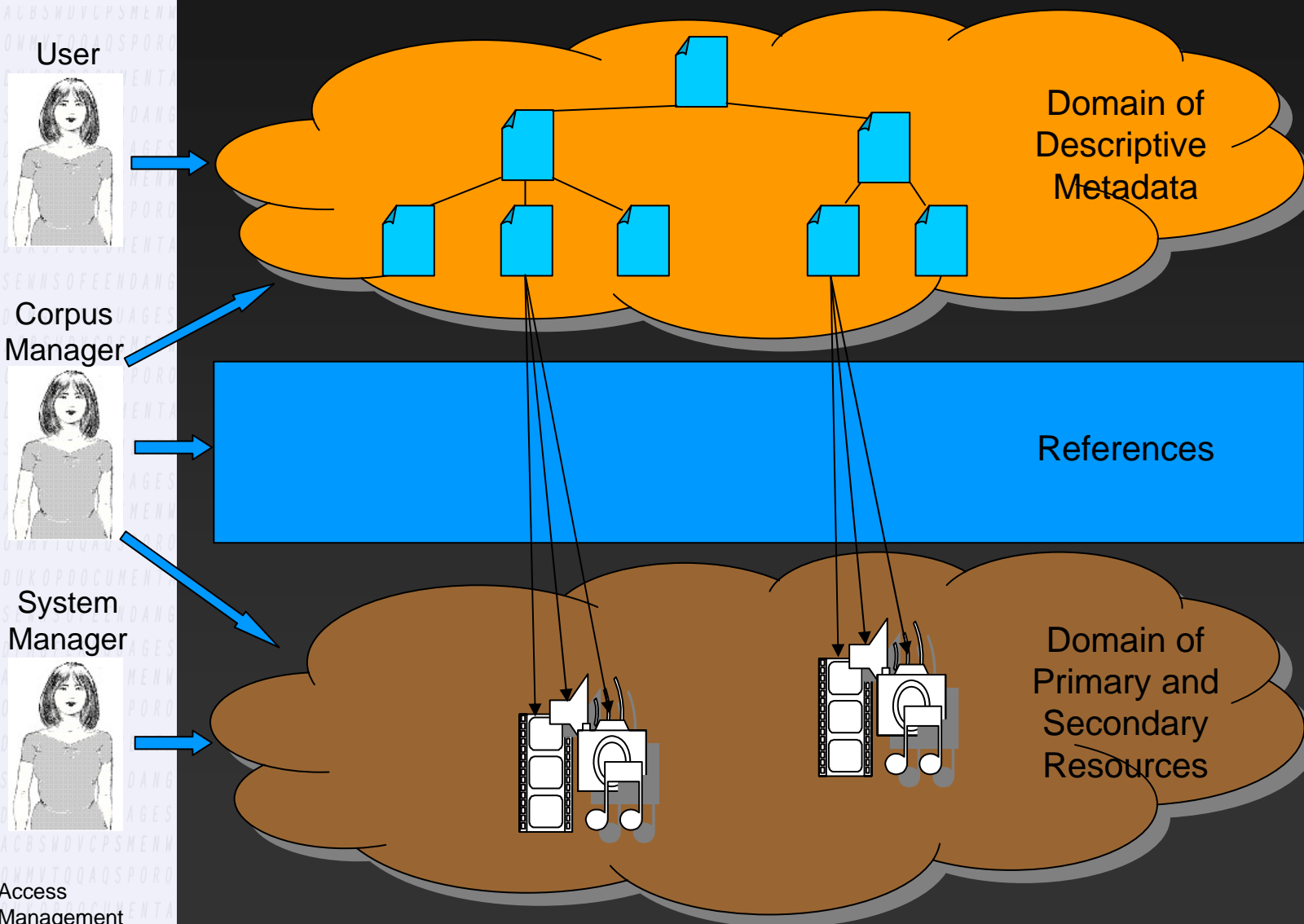
Idea is simple:

- assemble your own work space (WS) from the archive (annotated media, lexica, texts)
- do searches on MD and content on this WS
- compare several segments, entries, etc
- jump between lexicon, annotation and texts
- draw relations of different types
- add collaborative comments
- work currently in progress



The screenshot shows a web browser displaying a Chinese menu page. A comment box is overlaid on the page, containing the following text:

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



- open
- virtual
- linguistically ordered
- stable
- validated

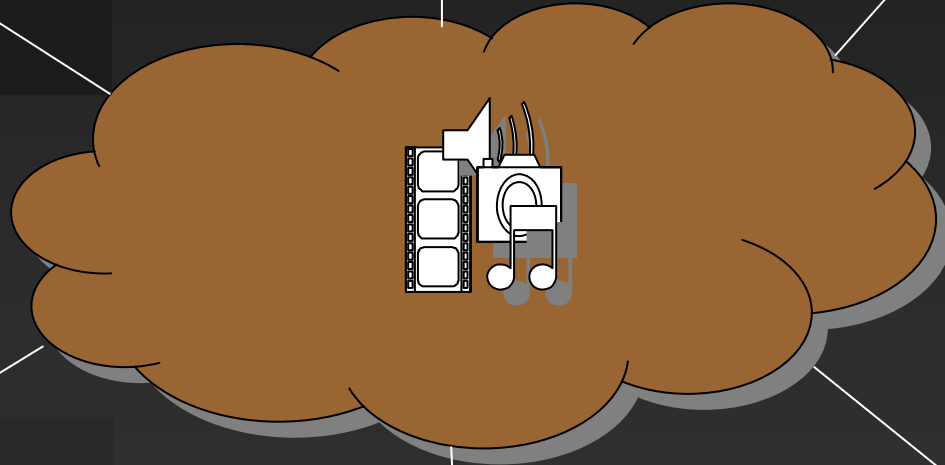
- checked
- URID level to come

- restricted
- physical
- technically ordered
- subject of changes

- dynamic copies to CC in Munich via AFS client
- push strategy
- protected channel
- yet pure backup

- each object is directly addressable (URL or dir path)
- no extra shell is needed
- no particular organization required

- resource copies to others



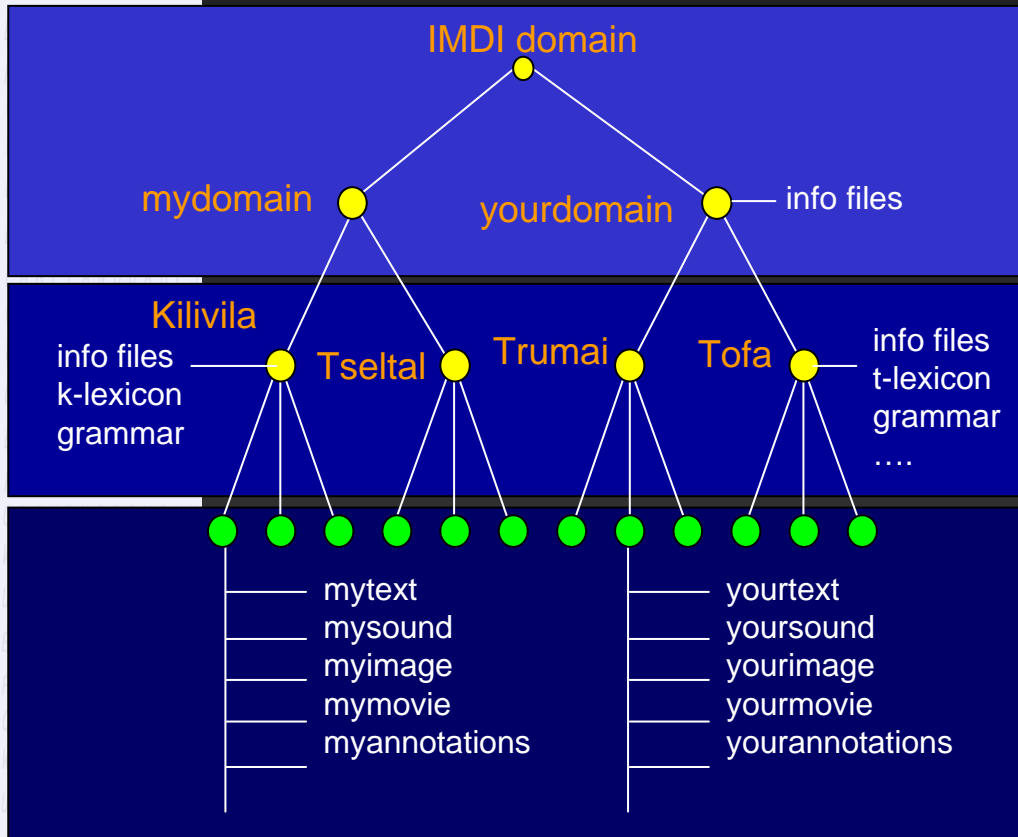
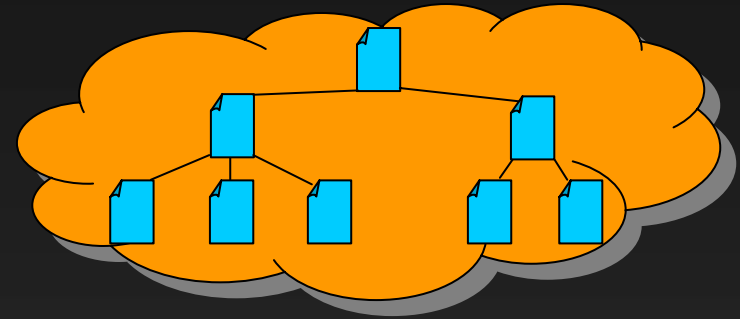
- dynamic copies to CC in Göttingen via RSYNC
- pull strategy
- yet pure backup

has a physical realization in a 3 layer HSM

- 3 copies automatically
- file type based strategies
- organization changes regularly

- complete copies to others

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

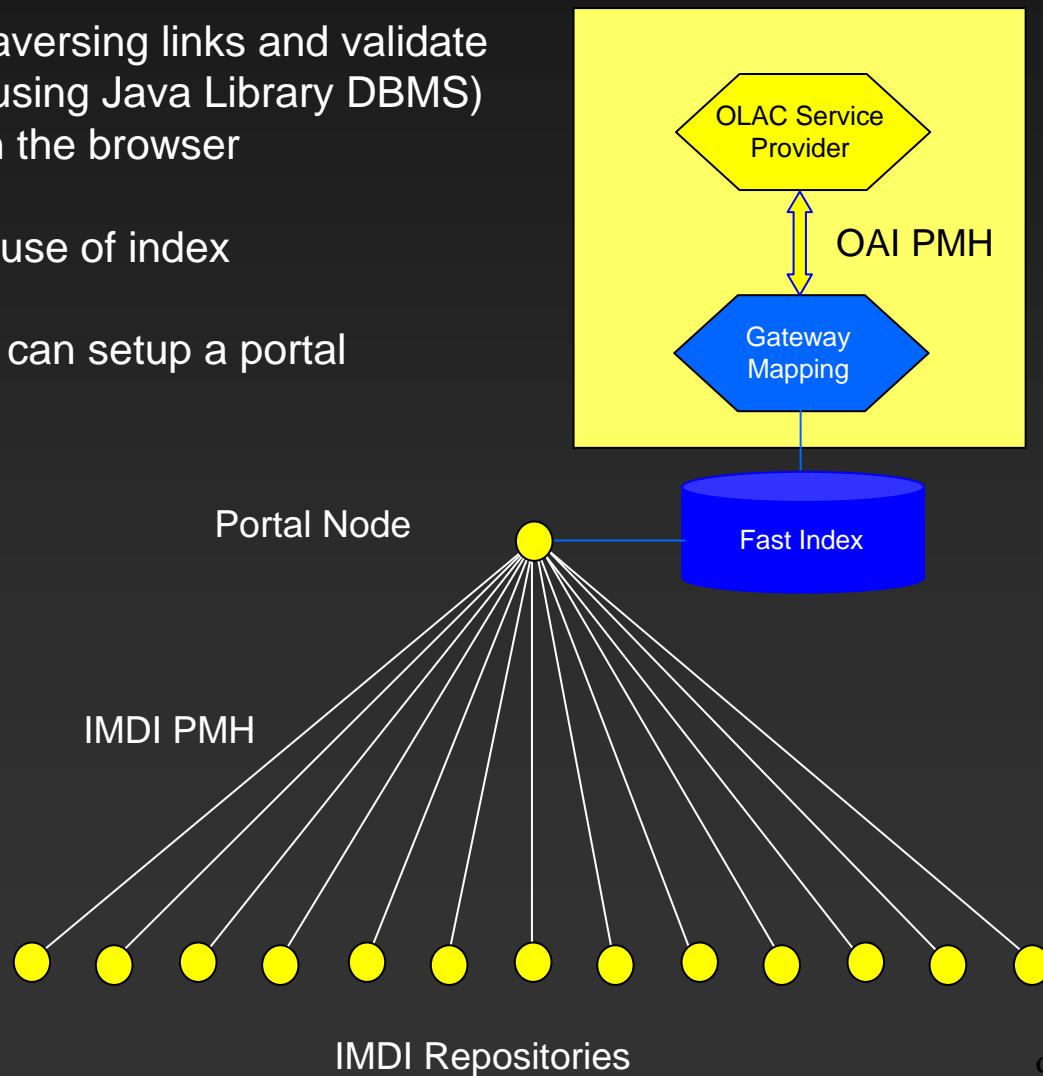


- fully distributed domain
- sufficient to register the root URL
- searching requires harvesting
- HTML browsing requires harvesting

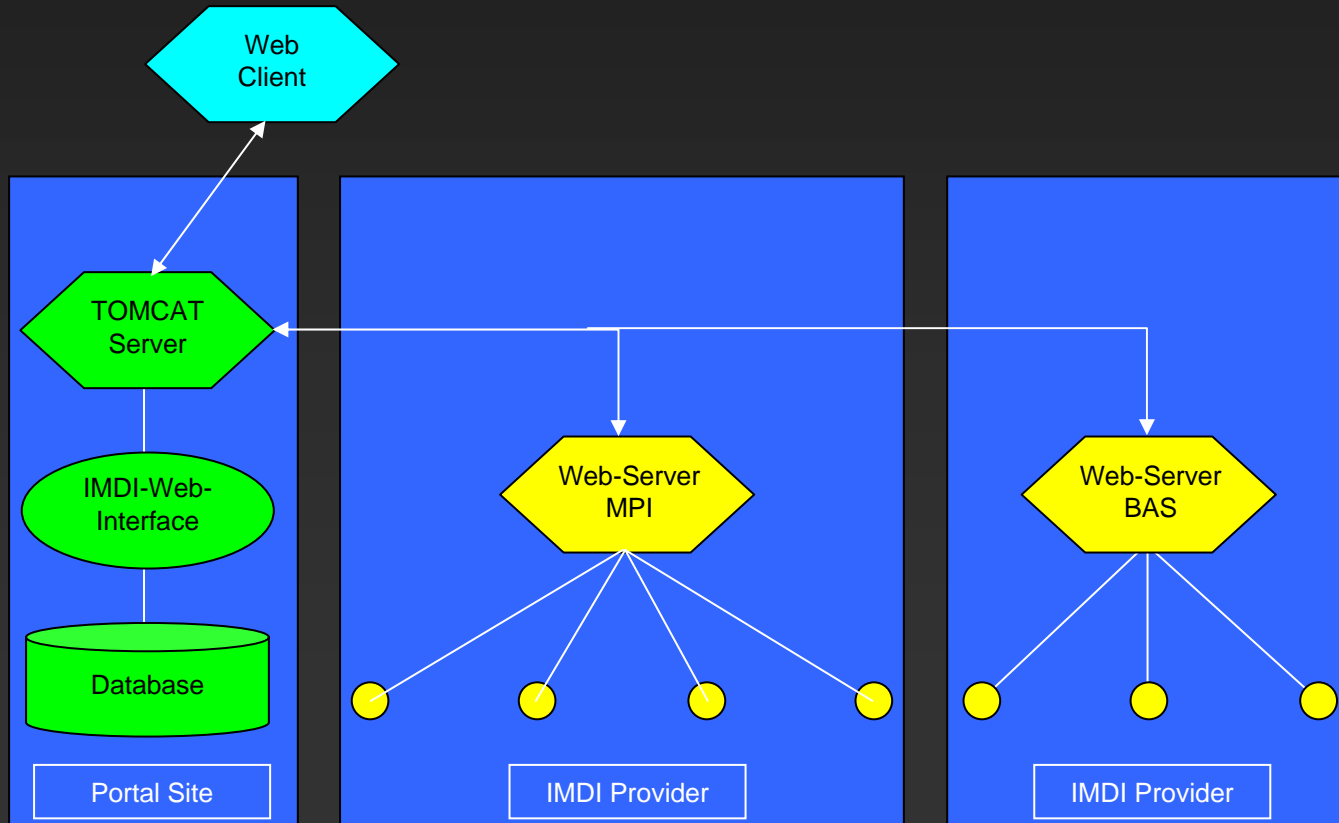
harvest all data by traversing links and validate
 create an index file (using Java Library DBMS)
 just select a button in the browser

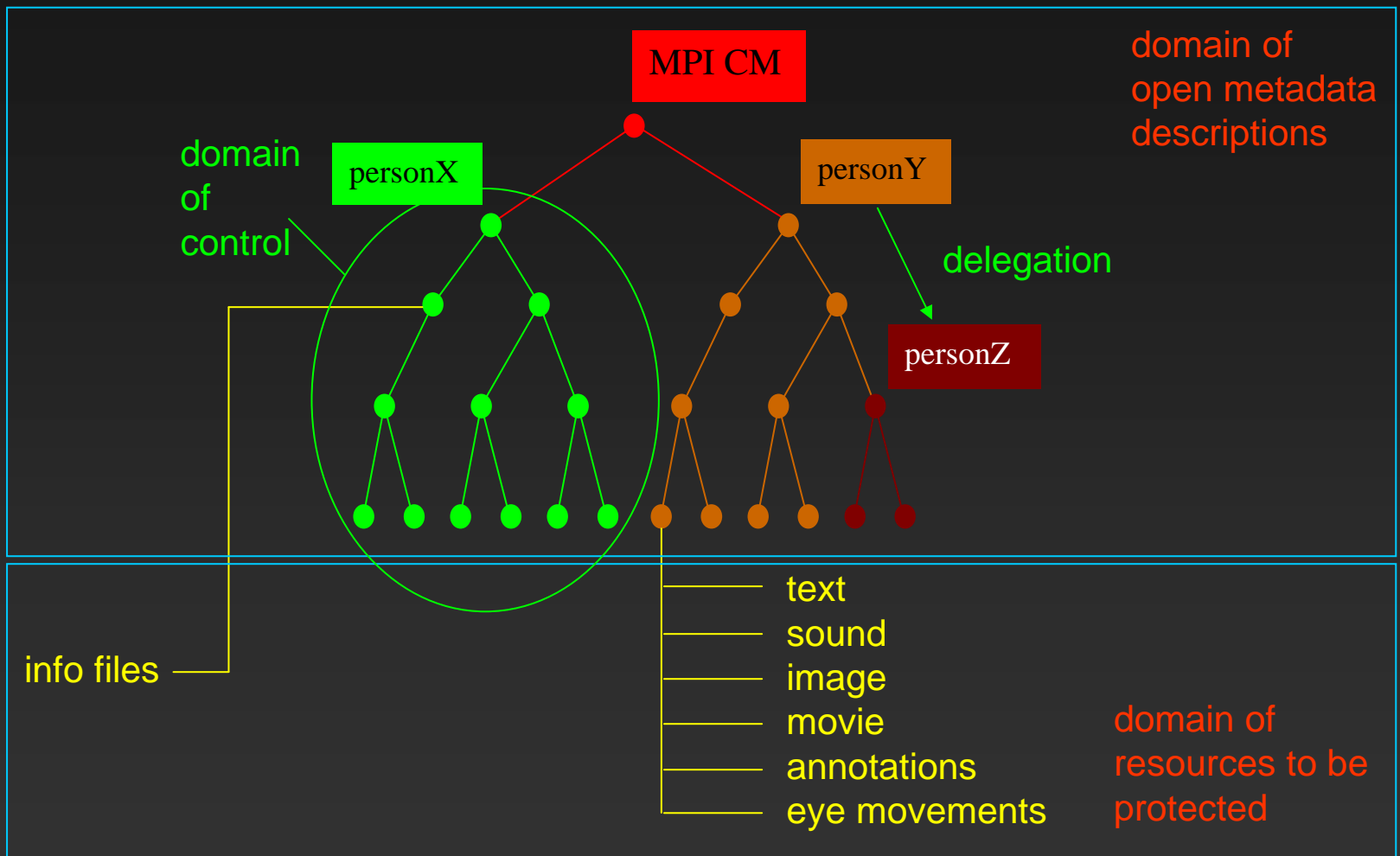
OLAC bridge makes use of index

so: simple, everyone can setup a portal

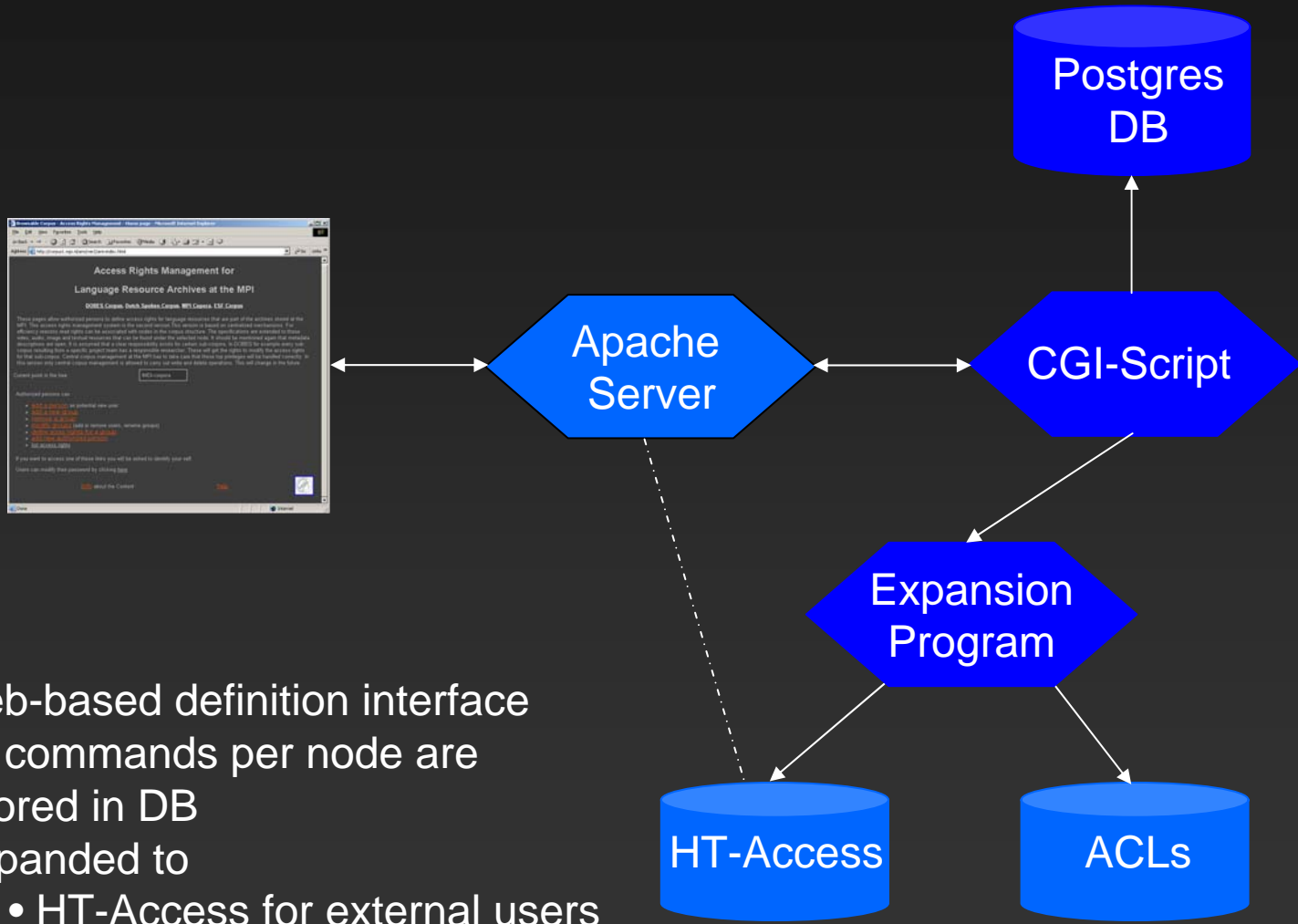


install Tomcat server and “IMDI-Web-Interface”
makes use of harvested metadata

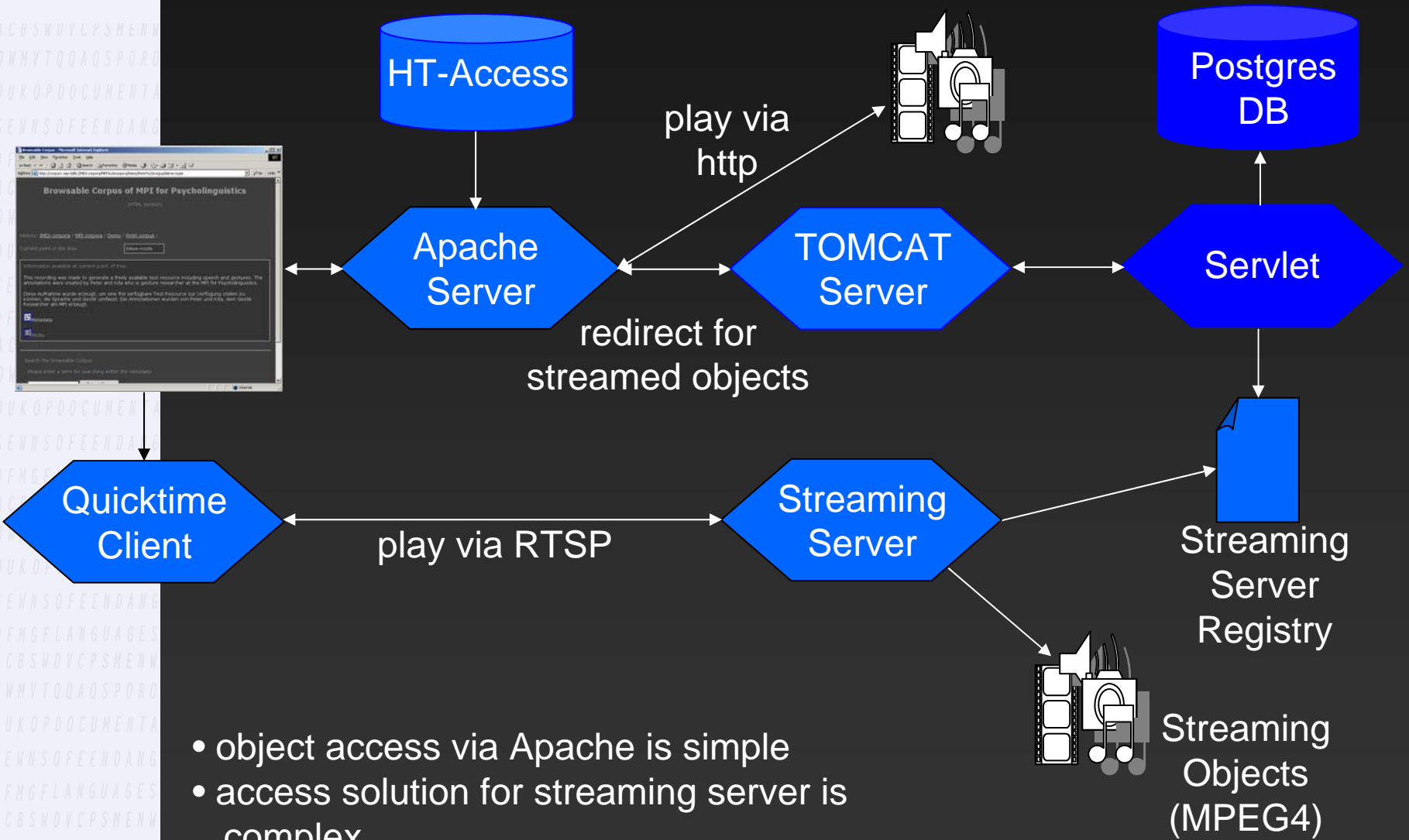




- current solution is centralized – one database
- has delegation mechanism to make administration tractable
- association of declarations etc is possible
- powerful commands from any node to give rights to groups



- web-based definition interface
- all commands per node are stored in DB
- expanded to
 - HT-Access for external users
 - ACLs for internal users



- object access via Apache is simple
- access solution for streaming server is complex



- what are the things we don't want to change resp. we need?
- adherence to a few agreed international “standards” – coherence
- every resource incl. metadata descriptions must be accessible without any additional shell (URLs / File System) can be additional shells
- IMDI as the catalogue system – at least for the time being
 - the core category set and its definitions (richness)
 - the capability of browsing
 - the capability of distributed operation
- after transition to URID system we have to stick to it
- principle of local operation (complete copy incl. MD)
- issue of long-term survival and long-term interpretability
- independency principle (fallback)
- very robust and stable services