

Perspectives for Accessing and Safeguarding Endangered Languages

the DOBES View

Peter Wittenburg
Max-Planck-Institute for Psycholinguistics

peter.wittenburg@mpi.nl
www.mpi.nl/DOBES



What is DOBES?

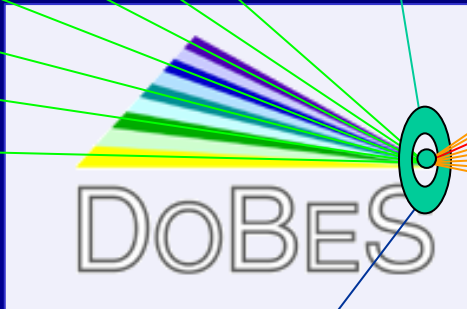
*presentation, education
(Books, CDROM, ...)*

Generations of Users:

- field researchers
- other researchers
- indigenous people
- students
- teachers/pupils
- journalists
- general public?
- ...

Potential Users
short-term impact
long-term impact

MPI Team



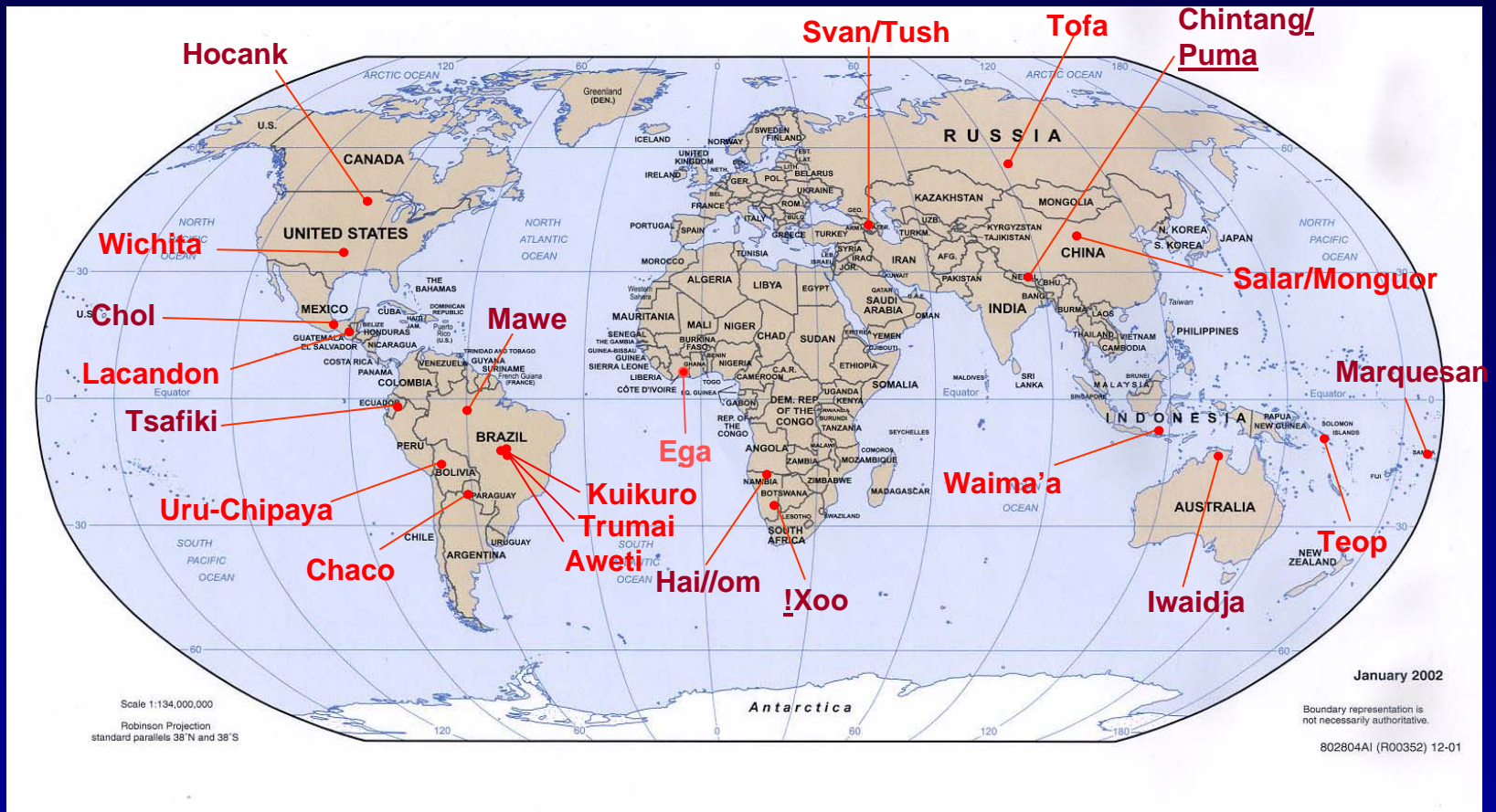
Archiving Team

VolkswagenFoundation Team

Documentation Teams
operating independently

- Waima'a
- Svan
- Lacandon
- Chaco Lang
- Uru-Chipaya
- Salar/Monguor
- Aweti
- Kuikuro
- Teop
- Tofa
- Trumai
- Wichita
- +9 others

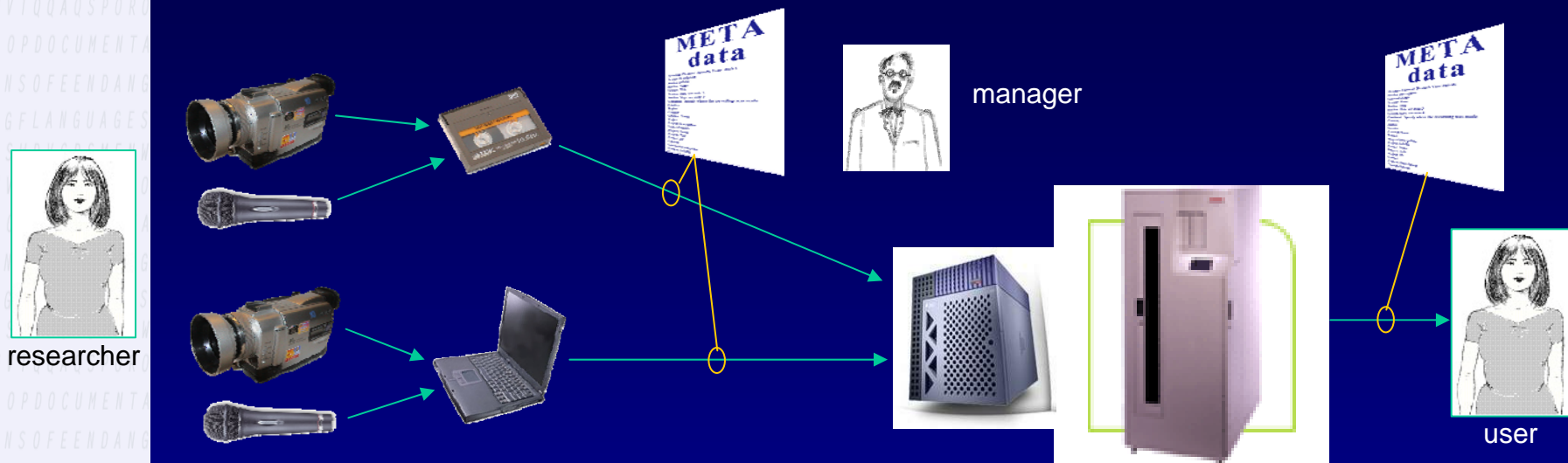
Languages in DOBES



- started September 2000 with 8 teams in a pilot phase
- after pilot phase extension to 12 teams
- currently 21 teams working in parallel
- expect two/three further calls (one per year)

What is the state?

Workflow utterly important - otherwise chaos



350 Digital Master Files (digital copies of a/v-tapes > 350 h)
 561 Sessions described by metadata and (therefore) organized
 49 Annotations, some first grammars and lexicons
 many Photos + field notes

at MPI much more (incl. EL, Child studies, Gesture studies, etc)
 total > 3000 h a/v > 15.000 sessions

all in a well-organized corpus based on the IMDI Metadata standard

Documentation Agreements

- language documentation to preserve human heritage
- intention to cover great typological difference in DOBES
- revitalization is recognized by all participants as another major goal
- language documentation capturing also its cultural background, facial expressions, lip movements, etc
- therefore focus on multimedia recordings and inclusion of ethnologists, musicologists, ...
- a few principles / agreements (bottom up approach)
 - teams have to give (copies of) their data to the archive
 - variety of text types and genres
 - 2 main tiers (orthographic/phonetic, translation in major language)
 - further glossing in local lingua franca and English
 - for some material deep linguistic analysis (could be Adv.Glossing)
 - good documentation of all tag sets (morphosyntactic terms)
 - provision of a lexicon (not “a” to “x”, but for example topic oriented)
 - sketch grammar, field notes, ...
- no agreements about tag sets etc (too complicated)

Technical Agreements

- help for archivist in organizing the material by IMDI metadata
- for the **archive** holding adherence to some standard formats such as XML, UNICODE, PCM 20kHz, MPEG2, MPEG1/4, JPEG **not MP3, MD and DV**
- for the **researcher** a recommendation to use tools that support the formats, but also support for

SHOEBOX
ELAN

nothing better
a mm/mm annotation and exploitation tool
creating/reading EAF (XML schema)

Transcriber
Praat
WORD
etc

an excellent tool for audio annotation
an excellent tool for sound analysis
have built a converter
some additional converters

Safeguarding Aspects

data survival primarily a matter of social acceptance (no control)

but there are aspects we can influence

three relevant layers:

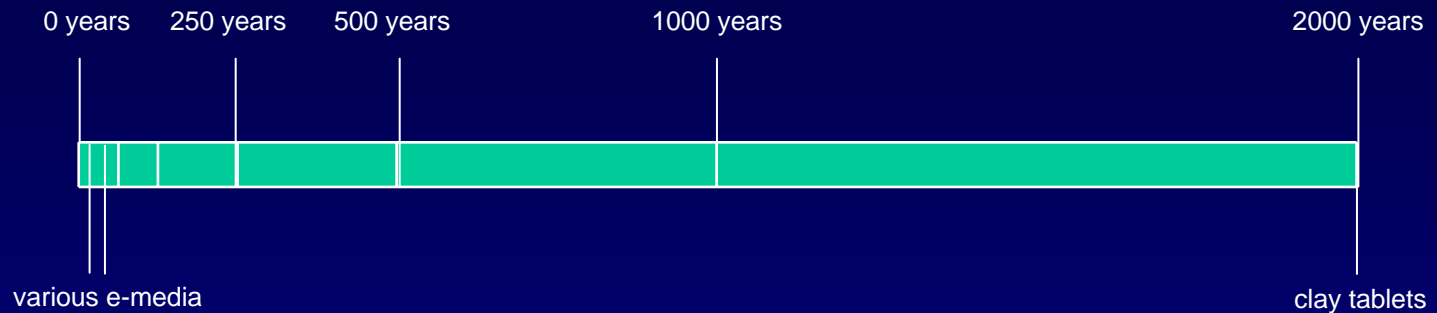
- physical storage layer
how long is data physically available?
- encoding layer
how long can data be interpreted?
- organizational layer
how long can we assure manageability?

may not forget:

speaking about dynamic online accessible archives

Physical Survival

figure indicates our media problem



persistence times

- hard discs ~ 4 years ?
- CDROMS ~ 10 years ?
- MPI institution ~ 20 years ??
- MPG society ~ 100 years ???
- Germany, Netherlands, ... ~ ?????????

- continuous migration
 - institute copying
 - campus copying
 - society copying
 - GRID copying
- replacement
 at MPI 2 copies, originals with scientists
 on campus 3. copy
 within MPG 4. and 5. copy at distinct places
 automatic, international copying

Physical Survival

- copying to compensate for small media lifetimes and errors
- distributing to compensate for political uncertainties

Consequences

- need to keep the costs low
- “hide” our data
- need efficient Data GRID technology
- need good ethical & legal agreements

Interpretability

011001010100001010110100101010



Guarantee interpretability of data independent of technology change

- there is no good solution - but hope
- need standards, education, awareness + understanding, discipline
- format and encoding migration (again a cost problem)

Is this really a long-term issue?

- some argue that future generations will be smart enough
- in addition: problems of decoding could increase attraction factor
- nevertheless: let's do our best now

Manageability

Guarantee coherent and accessible archive

- easily find the resource bundles
- easily discover interesting resources
- easily manipulate within the archive (add, move, copy parts, ...)
- DOBES (and MPI) is organized with IMDI Metadata
- DOBES applies the “immediate way”
do MD description/organization and conversion now

Is it really a long-term issue?

- not per se - although would be helpful
- needed for short/medium term usage
- how long will archive organization survive?
(new descriptors, new technologies, ...)
- nevertheless: let's do our best now, explicitness
- knowledge about archive content will decrease

Accessing the archive

A few relevant aspects to consider

- are creating a distributed and interconnected archive of EL
- will copy resources frequently
- users worldwide want to access the resources

What do we need?

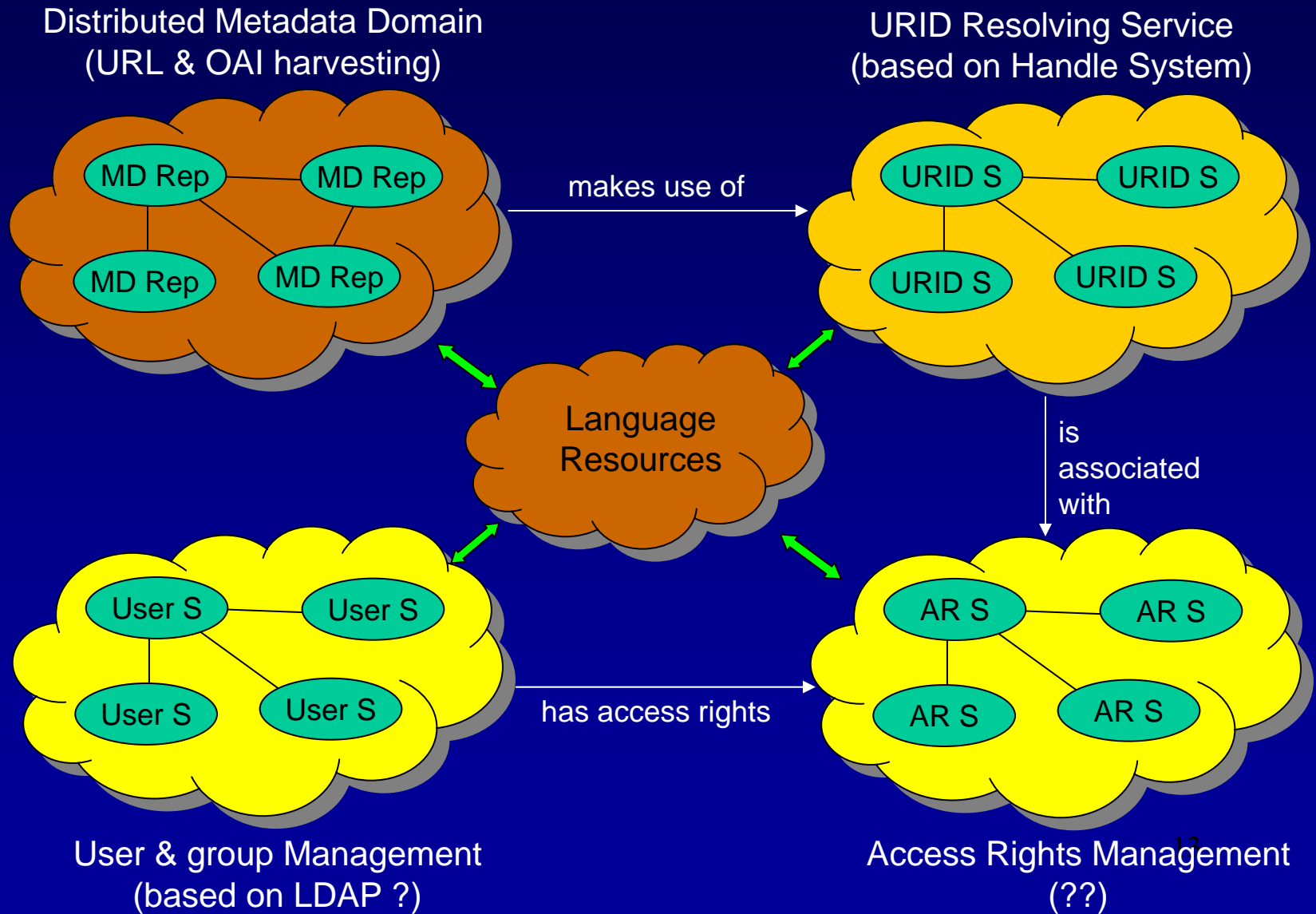
- a **fully distributed** system of services
- four closely related pillars
 - metadata to discover the resources
 - unique identifiers of all resources (paths not stable)
 - user administration
 - access rights management
- exploitation tools (not topic for today)

Should we do it alone?

- Noooo
- but we have to start implementing & testing
- it will be a complex scenario

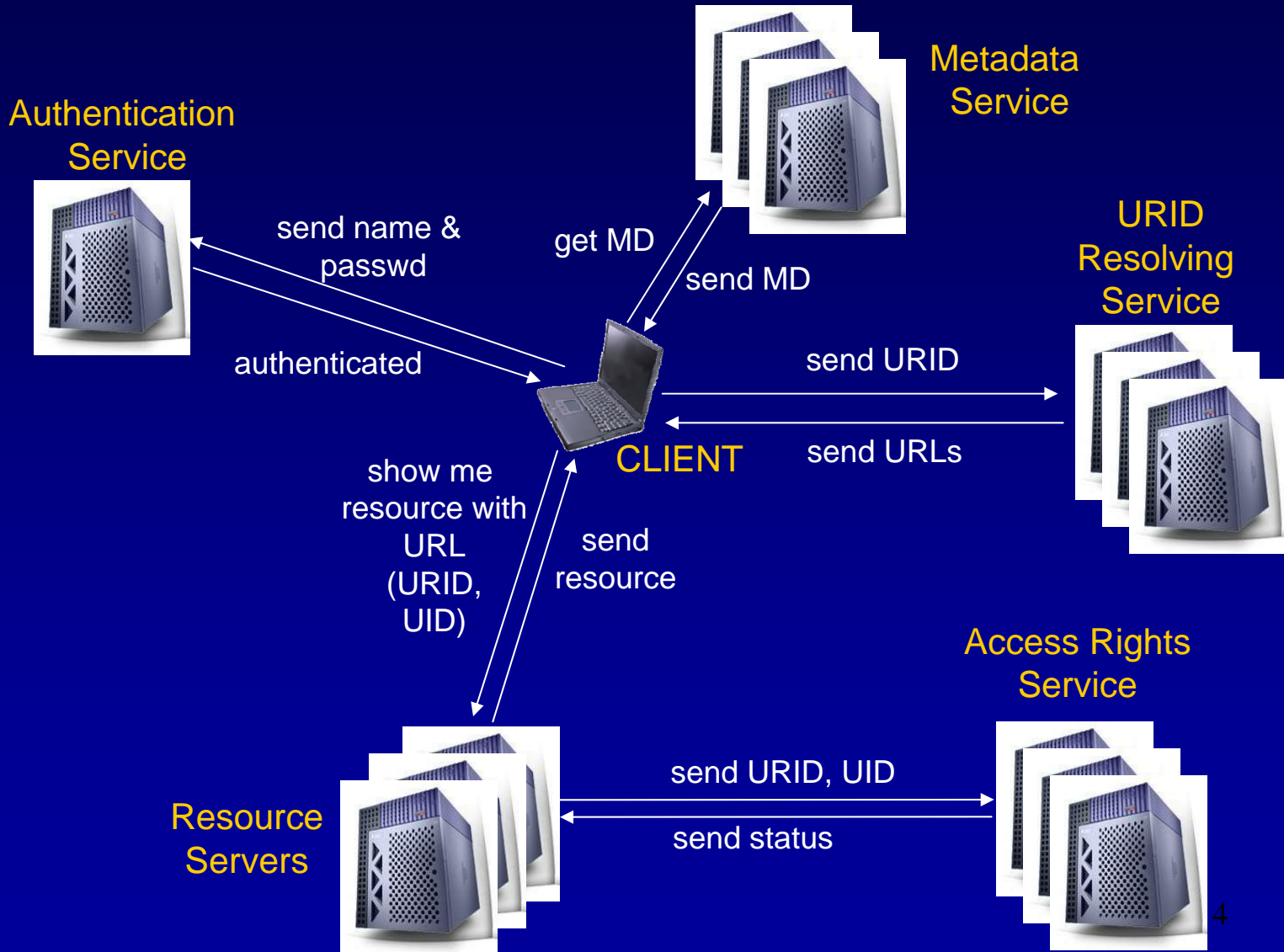
Resource M&A Service

Resource Management & Access Service (at MPI Nijmegen)



Operating Scenario

Different Scenarios - here one example (IMDI Browser)



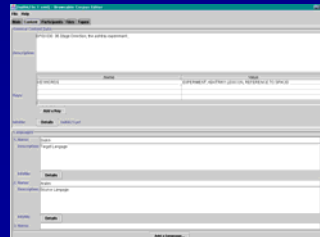
Metadata Infrastructure

All is based on the IMDI environment (set & tools)
some characteristics

- flexible structured set (mainly influenced by field linguists)
- all MD descriptions in open accessible XML files (XML schema)
- fully distributed framework (easy integration - just a URL)
- browsing in canonical trees provided by researchers for management
- support user specific tree building to create working domains
- automatic harvesting to support fast search
- when found a resource - immediate exploitation (start tools or copy)
- **two navigation shells: IMDI browser (Java) and HTML browse/search**
- editor supporting the set and controlled vocabularies
- bridge to DC/OLAC



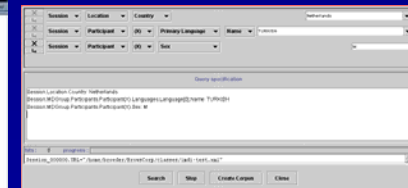
www.mpi.nl/ISLE



MD Editor



MD Browser



MD Search

www.mpi.nl/IMDI

URID resolving service

- no singular point of failure - distributed service
- hierarchy of naming authorities
- local naming conventions (flexibility)
- resolving URIDs from different environments (even standard html)
- need plug-ins and proxy services
- speed - scalability
- trust in support and quality of software

- chosen the Handle System
 - seems to offer almost all we want
 - questions as “what happens with parameters sent in http requests”

User and AR Management

- this is a very critical matter due to ethical and legal aspects
- so system must be maximally trustworthy
- no singular point of failure - distributed service
- delegation of managing users to many
- delegation of managing access rights
- speed - scalability
- trust in support and quality of software
- for user management and authorization a distributed LDAP solution is at hand
- for AR management nothing is so obvious yet

Conclusions

Working at new ways to offer EL resources.

- are approaching a situation where data can be made persistent by distributing them on a large scale
- this creates problems that require us to deploy a highly integrating system handling URIDs, users and access rights
- this is a complex system and it will cost a lot of time to get it running smoothly and reliably
- AND: this has to be developed jointly by several
- we need DELAN (Digital Endangered Languages Archive Network)



The End

Thank you