

Multimedia/Multimodal Language Corpora

Peter Wittenburg
pewi@mpi.nl

www.mpi.nl
www.mpi.nl/DOBES
www.mpi.nl/ISLE



Max Planck Institute for Psycholinguistics

What is the problem / why speak about this topic?

- many projects going on gathering valuable resources
(Swedish dialect corpus at Lund,)
- at our institute:
 - digitization and media linking of old material (ESF, P-Mol, ...)
 - about 20 field researchers + student assistants (staples of tapes)
 - lab recordings (GestureLab, ETLabs, SpeechLab, ChildLabs)
 - Endangered Languages Project (8 teams → 20 teams)
 - MUMIS project (multimedia indexing)
- **faced with a scaling up along several dimensions**
within institutions and especially across institutions



What are these dimensions?

- multimedia & all digital (many scientific reasons)
 - text only → text + video, sound, other time series
 - complex annotations (standoff format) →
 - number of related objects (organizational effort)
 - amount of storage space
(1h video = 1 GB MPEG1, 1h audio = 100 MB, 1h text = 100 KB)
 - new standards in dynamic world
- number of active researchers
- attitude towards language resources (not private capital)
- availability of language resources (standards, organization, ...)
- idea of long-term archiving has new dimension (no paper solution)



These are the points:

- single researcher can't solve the problems
- prevent an increasing CHAOS we already have it ☹
- within project: invest time & money on all these issues
define project strategy, adhere to common formats
- within institution: invest time & money on all these issues
define common strategy, adhere to common formats
- in the following: methods at the MPI (still fighting with details)



Basic Decisions

- early decision towards an **all-digital world**
 - Immediate access to the raw data, can't do av copying anymore (n times rt)
- early understanding that resources have to be **shared in Intranet/Internet**
- therefore working on **fully distributed infrastructures**
- **basic requirements for processing and tools:**
 - describe/register resources directly when created
 - let user operate in a conceptual domain (requires elaborate descriptions)
 - hide physical structure and make access format & platform independent
 - access should be independent of location (next generation mobile computing)
 - allow network-wide collaboration

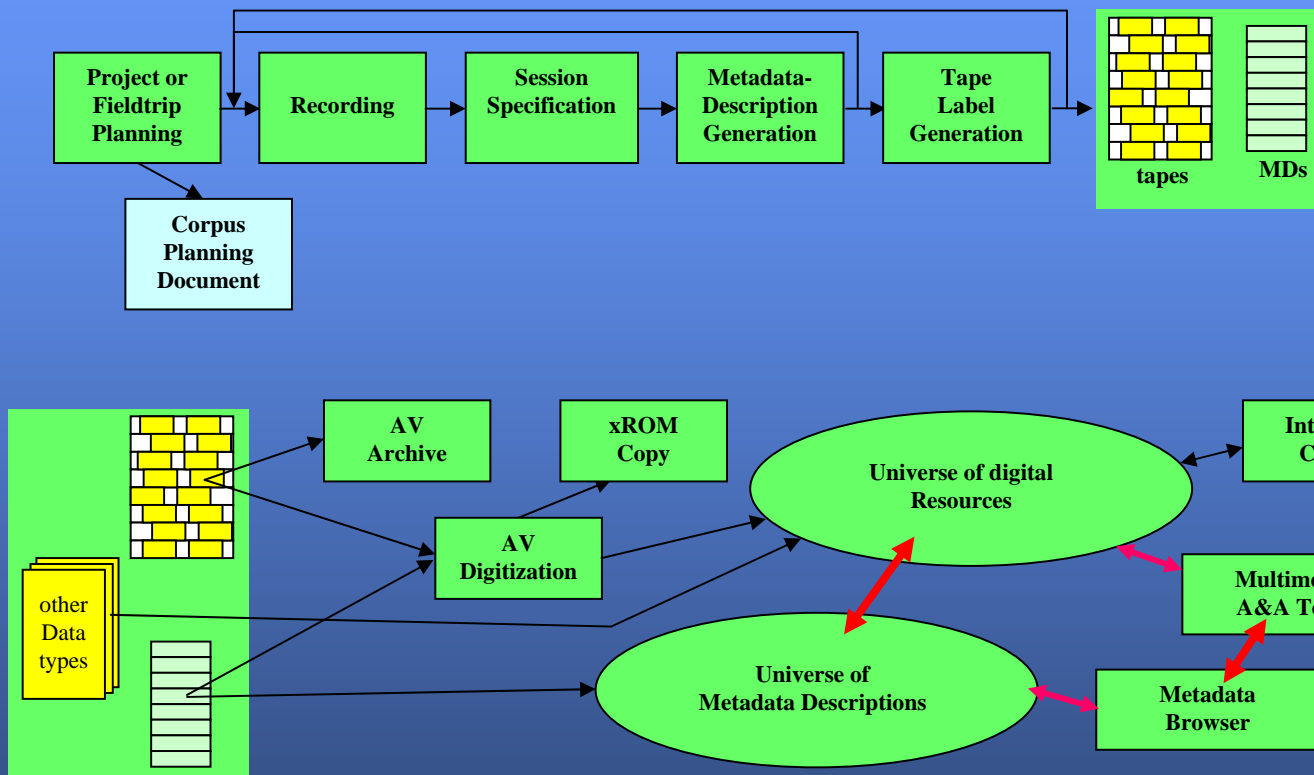


Key Elements of Infrastructure

1. workflow scheme for data processing (machinery)
2. redundant & proven infrastructure for digitization & compression
3. structured repository of metadata descriptions (find & execute)
4. format-independent, Internet-capable Multimedia A&A tool
5. reliable archive infrastructure



Workflow Scheme



creation process

- “t-exact” devices
- digitize later/first
- time code aspect

archiving process

- split audio/video
- archive manager



Key Elements of Infrastructure

1. workflow scheme for data processing (machinery)
2. redundant & proven infrastructure for digitization & compression
3. structured repository of metadata descriptions (find & execute)
4. format-independent, Internet-capable Multimedia A&A tool
5. reliable archive infrastructure



Digitization & Compression Infrastructure

- all-digital world
- define small **range of equipment** (DV, DAT, MD, CR)
but also VHS, Hi-8, VCD, Uher 4400, ...
- give **guidelines** how to do recordings (continuous mode, gaps, ...)
- rely on **open** (de facto) **standards** (MPEG1/2, wav)
what about MD and MP3 compression??? (MP3: 128 kbps - HiFi norm)
- test **synchronization correctness**
- test **software compatibility** (QT, JMF, ...)
- **technology & user driven** >>> MPEG2 (3-6 Mbps vbr)
why not DV or MPEG I-frame only (>factor 10)???
- redundant & efficient **digitization setups** (4 video, 2 DAT audio)
- many **scripts** (conversion, split, integration)



Key Elements of Infrastructure

1. workflow scheme for data processing (machinery)
2. redundant & proven infrastructure for digitization & compression
3. structured repository of metadata descriptions (find & execute)
4. format-independent, Internet-capable Multimedia A&A tool
5. reliable archive infrastructure



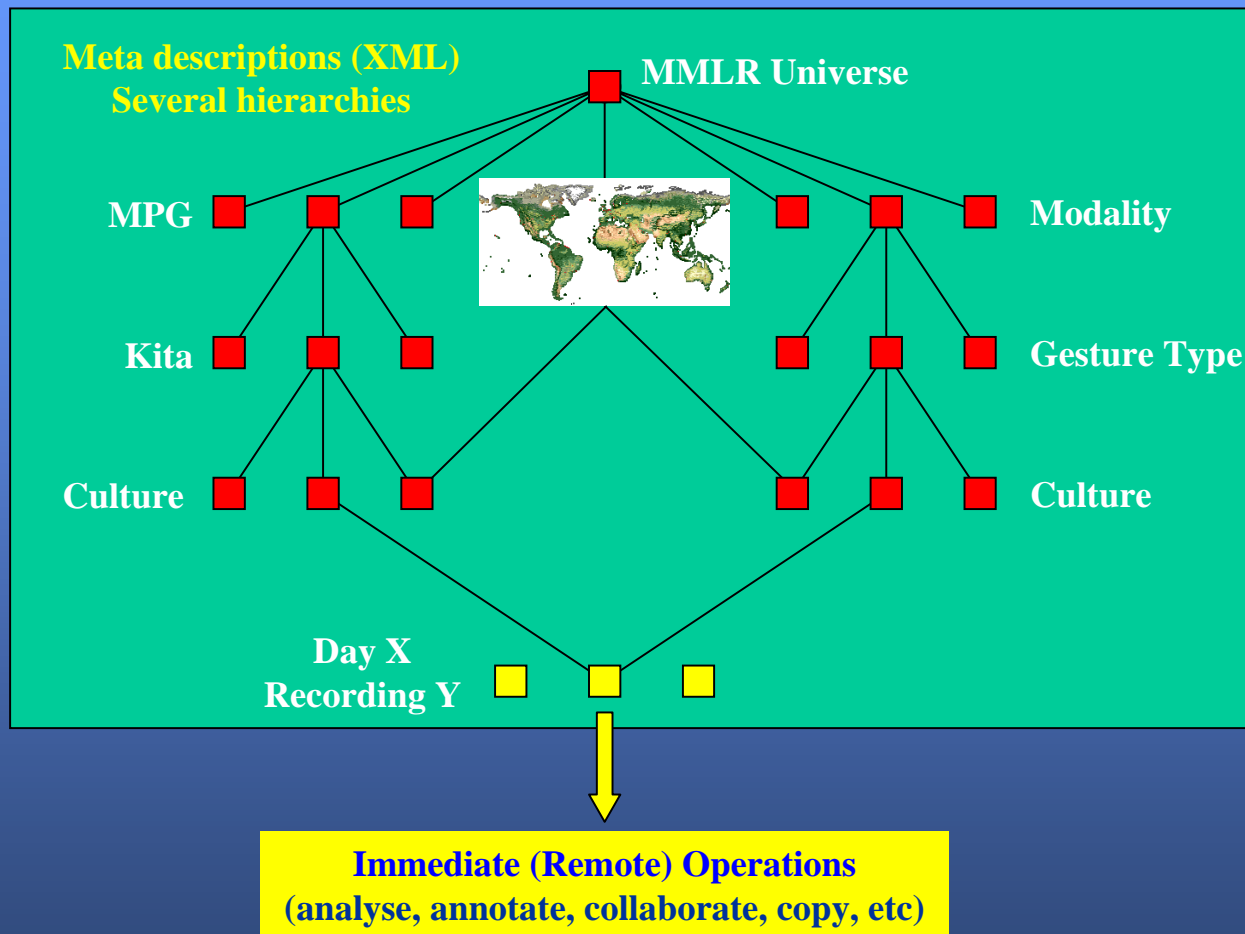
Metadata Descriptions

- **MD** is data about data – here limited element set to characterize language resources
goal is to find relevant resources and to have a quick check
- **MD universe is browse&search domain primarily for community members**
But also for general public
- **MD set has to be rich for community members!**
“Give me all LR with female speech in Yaminyung where gestures are coded”

- provide **MD editor** in the field – automatic creation of XML files
- provide **web-accessible linked hierarchies** of MD
- provide **web-capable MD browser**
- MD description are **open** – resources often not, nevertheless find them
- rely on **MD standard**: syntax-XML, structure-RDF/Schema
 - accepted semantics
 - MPI: early BC >>> IMDI proposal



Future Access Scenario to LR



- web-info basically not accessible
- easy navigation in metadata subspace
- distributed metadata resources
- browsing & searching & immediate operation

demo

- world wide activities
 - W3C
 - MPEG7
 - DC
 - OAI
 - ISLE (MPI)
 - ...



MPI Metadata Editor

File Edit Help Structure

Main Data

Name: liela17k.1

Description: This is the ESF subcorpus TL English, SL Italian, subject Lavinia, cycle 1, sequence 7

Access: Free

Project Controller: ESF

Keys:

Name:	Value:
SOUND_LINKING	??

Content Participants Files

Transcription Files

Name:	Format:	Remark:	Browse:
1. /data/corpora/esf_conv/data/english/longitudinal/italian/liela/liela17k.1.cha	text/chat	Was generated from ESF transcript	Browse
2. /data/corpora/esf_conv/data/english/longitudinal/italian/liela/liela17k.1tr	text/esf	Original ESF transcript	Browse
3.	unknown		Browse

Media Files

Name:	Format:	Start:	Duration:	Quality:	Remark:	Browse:
1. /data/corpora/esf_conv/data/english/longitudinal/italian/liela/liela17k.1.sd	audio/esps	unknown	unknown			Browse
2.	unknown					Browse
3.	unknown					Browse

Label Files

Name:	Format:	Start:	Duration:	Remark:	Browse:
1.	unknown				Browse
2.	unknown				Browse
3.	unknown				Browse

Add a transcription file... **Add a media file...** **Add a label file...**



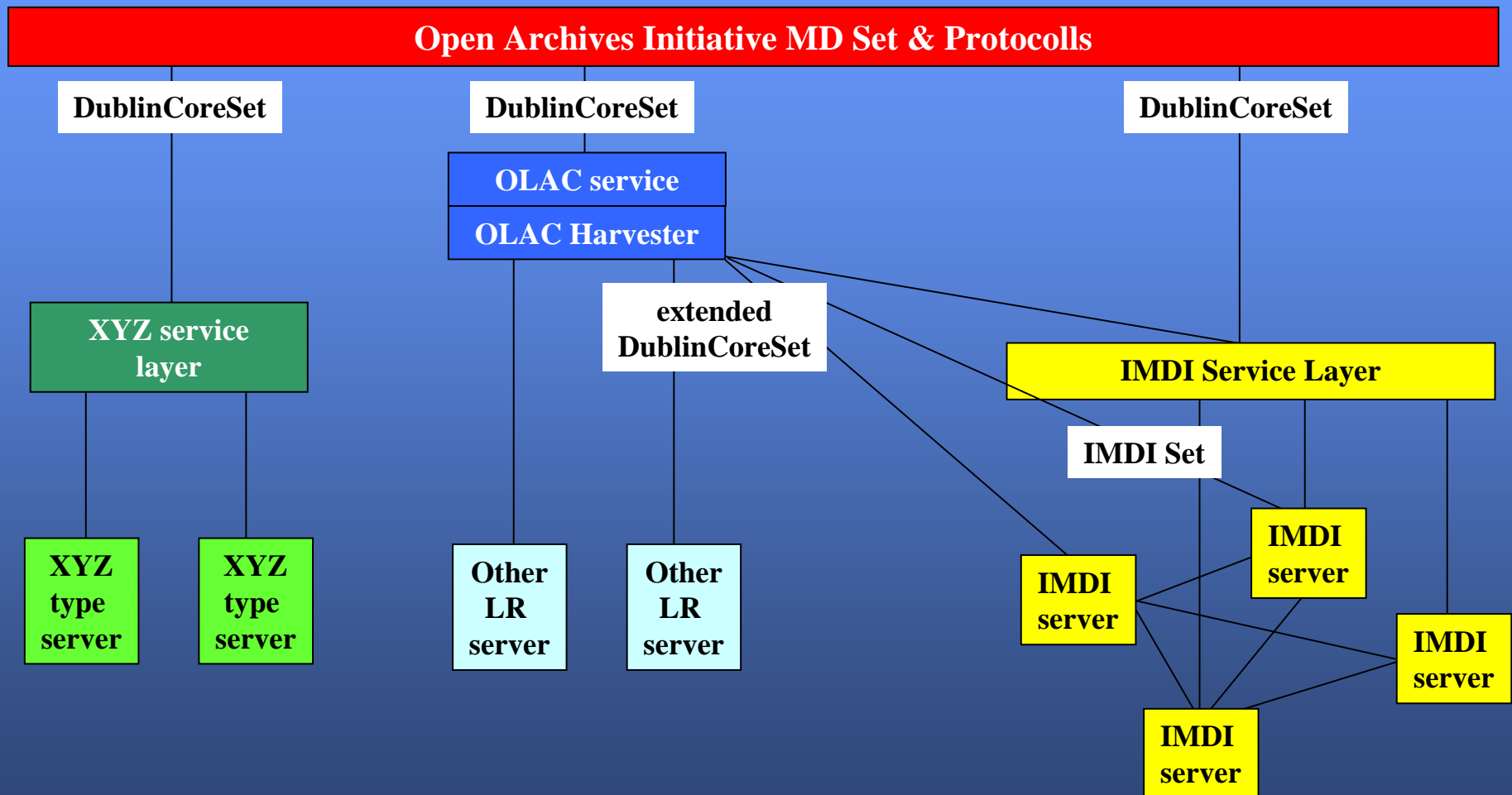
IMDI Metadata Browser

The screenshot shows the Netscape browser window with the following components:

- Browser Action:** Exit, About, Options buttons.
- History:** MPI Corpora (HTTP), file:d:/users/broeder/corpo
- Meta Descriptions:** A tree view showing a hierarchy: LAC-corpus Language & Cognition > africa > australia > arrente > guugu yimithirr > jaminjung > Schultze-Berndt > Field methods (selected).
- Root URL:** file:d:/users/broeder/corpora/lac/xml/lac.xml
- Description:** This file gives information about the fields method and about the types of data
- Info/Content:** Field methods and Types of Data. Linguistic documentation and description of Jaminjung and Ngaliwuru is comparatively scarce (see "References"). The materials presented here all come from my own fieldwork [may be modified], undertaken between April 1993 and July 1997, and extending for 22 months in total. During the first field trip in 1993, I lived in Bulla Camp for seven months. In subsequent field trips, I did not permanently live with Jaminjung and Ngaliwuru speakers or their families. However, I attempted to spend as much time as possible, under these circumstances, with speakers and other people with whom I had developed a personal relationship. This type of participant observation involved overnight visits to people living in outstation overnight bush trips, shorter fishing and hunting trips, or - more rarely - visits to specific significant sites. It also involved joining people in their Communities or in public places during



Global MD Perspective



MD - DublinCore

- 15 Elemente im Standard / Qualifiers in Diskussion (4 Jahre!)

Title	name of the resource
Creator	entity primarily responsible for content
Subject	topic of the content (OLAC: language written about)
Description	account of the content
Publisher	entity responsible for availability
Contributor	entity responsible for making contributions to content
Date	event in life cycle
Type	genre of content
Format	physical or digital manifestation
Identifier	unambiguous reference
Source	reference to a resource from which resource is derived
Language	language of intellectual content (language written in)
Coverage	extent or scope of the content
Rights	IPR information



Key Elements of Infrastructure

1. workflow scheme for data processing (machinery)
2. redundant & proven infrastructure for digitization & compression
3. structured repository of metadata descriptions (find & execute)
4. format-independent, Internet-capable Multimedia A&A tool
5. reliable archive infrastructure

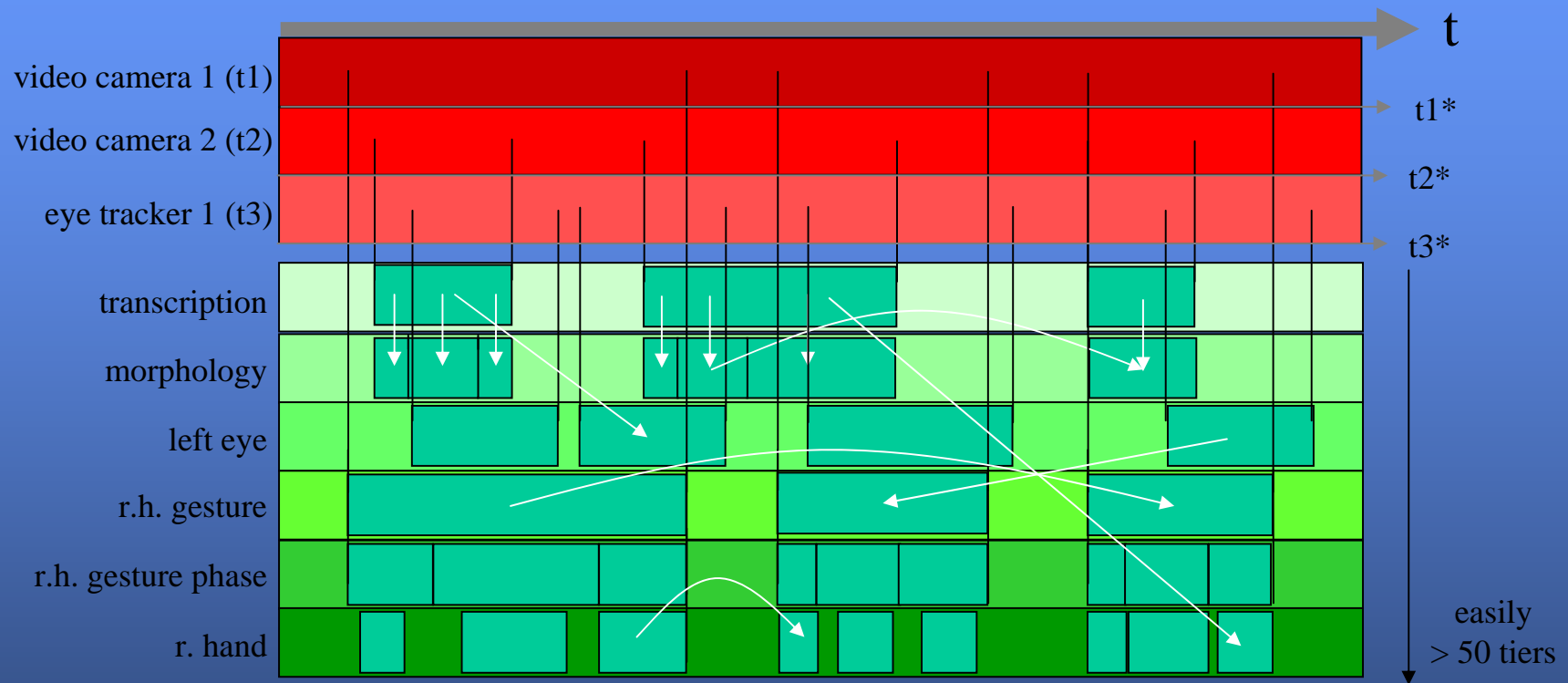


EUDICO: Multimedia A&A Tool Set

- platform-independence (Java – but QT on MACs)
- format-independence (Abstract Corpus Model as internal basis)
 - I/O modules for CHAT, Shoebox, rDB, XML (support old formats)
- several synchronized text/sound/video viewers
- integrated search tool
- media streaming via networks i.e. required fragments only
- distributed and local operation
- integration of professional sound analysis (PRAAT)
- type definition and data entry tool (type and structure of annotation tiers)
- UNICODE support
- improved generic search tool
- multimedia lexicon
- generic XML support >>> Atlas Interchange Format



Complexity of MM/MM LR

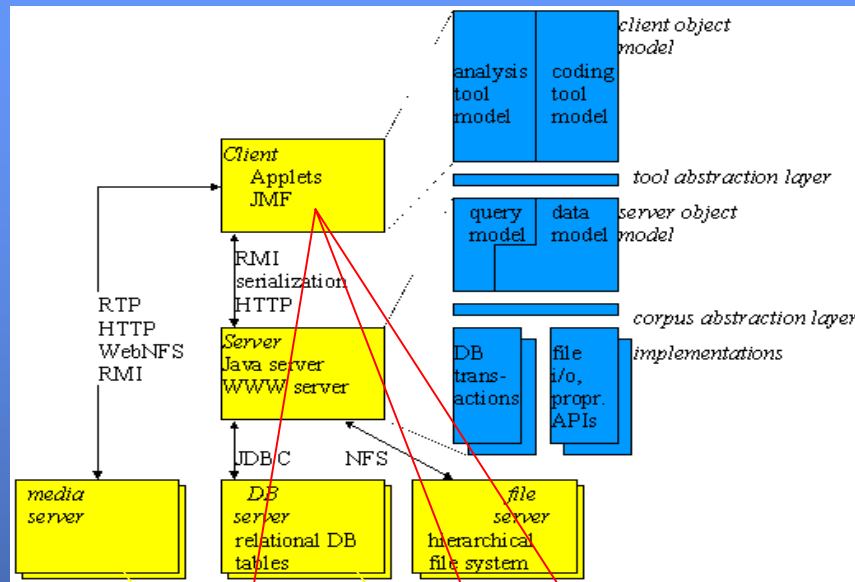


time scales, independent streams, partial time alignment, large nr. tiers, hierarchies, (labeled) references



Max Planck Institute for Psycholinguistics

Future Collaborative Scenario



SpchTr	Transl	Time Interval
60	ka bin k'	00:01:00.560 - 00:01:01.720
61	ti' ump'el tankabil chak'an beya'	00:01:01.760 - 00:01:03.560
62	pero tu'x'bil tun chak'an beya',	00:01:03.600 - 00:01:06.440
63	naach tun bin	00:01:06.480 - 00:01:08.200
64	bey te' lak'in beya',	00:01:08.240 - 00:01:09.120
65	ku yilik bine'	00:01:09.160 - 00:01:09.800

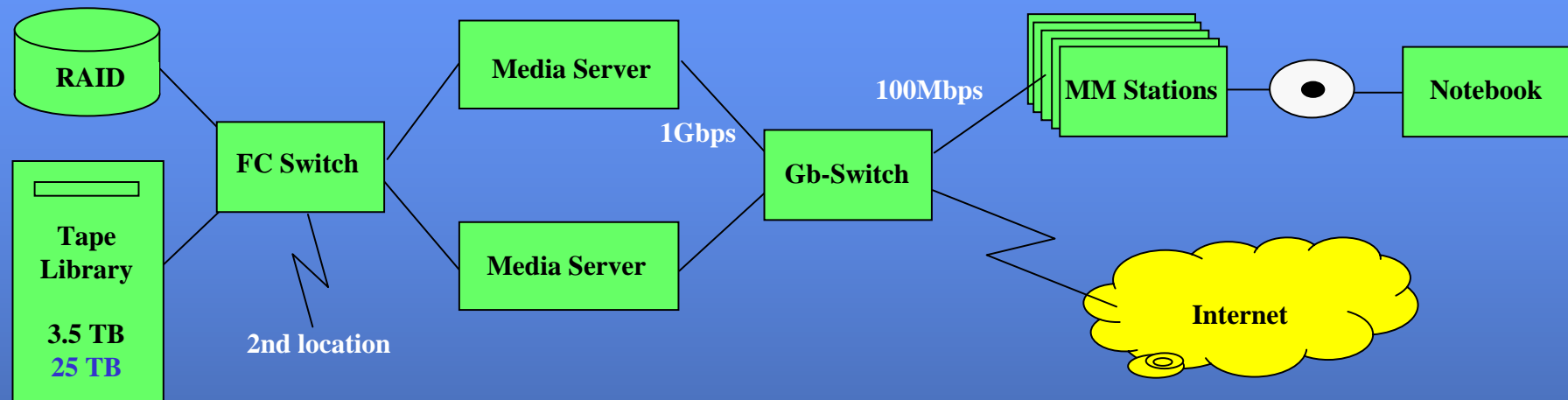


Key Elements of Infrastructure

1. workflow scheme for data processing (machinery)
2. redundant & proven infrastructure for digitization & compression
3. structured repository of metadata descriptions (find & execute)
4. format-independent, Internet-capable Multimedia A&A tool
5. reliable archive infrastructure



Archive Infrastructure



- efficient & reliable **storage management**
(near-line capacity, media change, 2. Location)
- **high storage capacity** (now 3.5 TB, in short 25 TB)
- powerful **media servers**
- powerful **network**



Conclusions

- have found answers for the most urgent problems (chaos prevention)
- costs much support time at this moment
- have a few tools ready (MD) and in preparation (EUDICO) which may help
- hope on some funds to create a European (+) corpus infrastructure
- are acting in a dynamically changing world – current solutions obsolete in few years?
perhaps all wrong what we are doing?
- working hard ☹
- all software free for academic usage ☺

