

# Web-based Language Documentation & Description at the MPI and within DOBES

Peter Wittenburg, Hennie Brugman, Daan Broeder  
Max-Planck-Institute for Psycholinguistics  
[pewi@mpi.nl](mailto:pewi@mpi.nl)

[www.mpi.nl](http://www.mpi.nl)  
[www.mpi.nl/DOBES](http://www.mpi.nl/DOBES)  
[www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)



Max Planck Institute for Psycholinguistics



**Web-based:** distributed services, world-wide accessibility, easiness of access  
all-digital, open availability, based on open standards, multilingual

→ **Description:** language & cultural background!, multimedia recording,  
linguistically neutral, extendable (complete)

### **Documentation:**

long-term availability, validated, ethically correct,  
responsible, person-independent

### **Archive & Data Infrastructure:**

- archiving is about long-term housing (multimedia) documents
- data is about sharable objects describing languages
- infrastructure is about mechanisms to setup, link and access such archives



# Background of work

- early decision towards an **all-digital world**  
driven by cognition & gesture research
- **emerging chaos** due to about 30 field researchers
- early understanding that resources have to be **shared in Intranet**
- therefore: since 3 years working on **distributed infrastructure**
- since August **DOBES project** (Volkswagenstiftung, now 8 - later >20 teams)
  
- **basic ideas:**
  - describe/register resources directly when created
  - let user operate in a conceptual domain (requires elaborate descriptions)
  - hide physical structure and make access format independent
  - access should be independent of location (next generation mobile computing)
  - allow network-wide collaboration

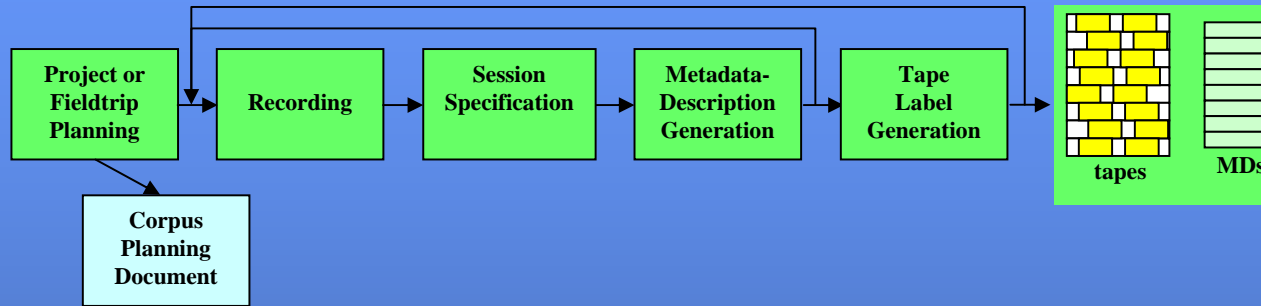


# Key Elements of Infrastructure

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

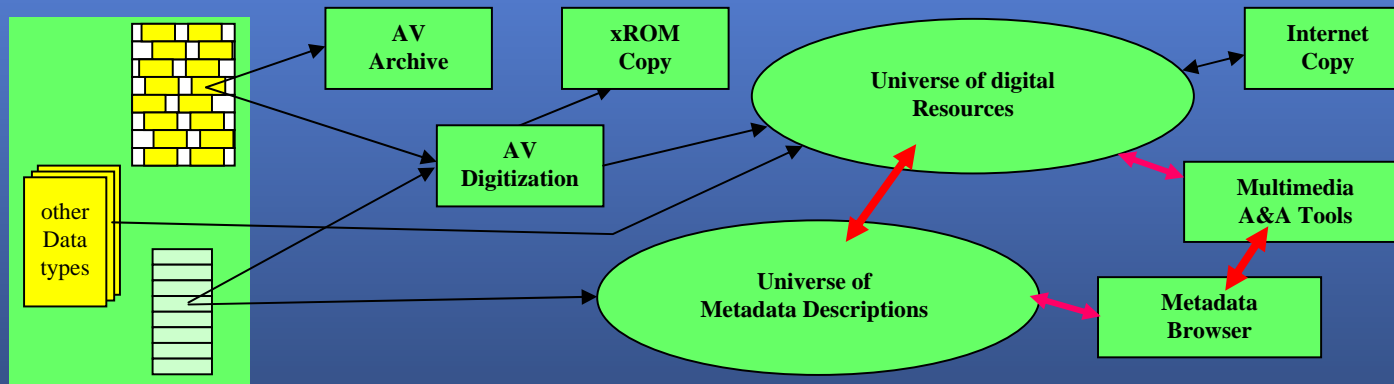


# Workflow Scheme



## creation process

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



## archiving process

- split audio/video
- archive manager



# Digitization & Compression Infrastructure

- all-digital world
- define small **range of equipment** (DV, DAT, MD, CR)  
in DOBES also 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 setups and **scripts** (conversion, split, integration)
  
- **DOBES:**  
wav – 44.1 kHz/48kHz  
MPEG1&2



# Metadata Descriptions



- MD universe is browse&search domain for community members not for general public!

- MD set has to be rich!

*“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



- rely on MD standard: syntax-XML, structure-RDF/Schema
  - semantics
  - MPI: early BC, DOBES: preliminary Core Set >>> IMDI proposal
  - scripts for conversion

talk tomorrow

- integrate & map into OAIMS for general public



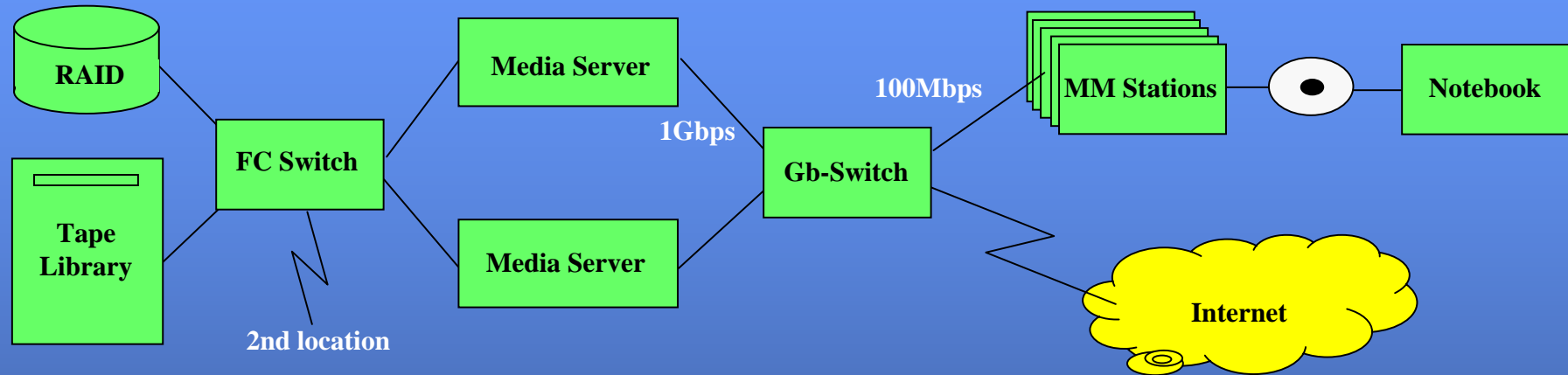
# EUDICO: Multimedia A&A Tool Set

- **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**

tool



# Archive Infrastructure



- efficient & reliable **storage management**  
(near-line capacity, media change, 2. Location)
- **high storage capacity** (n TB, 1 h MPEG1 = 1 GB)
- powerful **media servers**
- powerful **network**



# DOBES Data Types and Formats

- **Genres** (extendable list)  
common names
- **Media Data** (av, photos, pitch contours, laryngograph, ...)  
common abbreviations
- **Annotations Tiers** (mandatory set, but extendible, user definable)  
common names
- **Data Types** (mm corpora, lexica, MD, wordlists, grammar notes, ...)  
common abbreviations, common formats
- **Font Sets** until now (IPA, Latin, Chinese)  
Unicode

details



# MPI/DOBES/ISLE Contribution

Data Type	Data Models & Archives	Data Creation		Data Access	
	Store	Create	Convert	Display	Query
Metadata	ISLE-IMDI MetaSet MPI MetaSet DOBES&MPI MD	XML-MD Editor	scripts	MD Browser	MD Browser
Word List	<b>DOBES</b>				
Lexicon	<b>ALM</b>	EUDICO	?	EUDICO	EUDICO
Annotated Signal	<b>DOBES</b> <b>MPI</b>	EUDICO MediaTagger	I/O components, scripts	EUDICO MediaTagger	EUDICO SEARCH, MT
Writing System					
Interlinear Text	<b>DOBES/MPI</b>	EUDICO		EUDICO	EUDICO SEARCH
Paradigm	?	?	?	?	?
Field Notes	<b>DOBES/MPI</b>			MD Browser	MD Browser
Description	?	?	?	?	?
Others	<b>Diverse</b>	Knowhow, scripts	scripts	?	?



# Final Remarks

- MPI supports idea of open interoperable web-based infrastructures
- many obstacles & continuously new ones (technological, access rights, ethic, ...)
- need steps to overcome them again and again
- maximal benefit of the LR community is primary goal  
general public is secondary goal
- need well-balanced approaches
  - 1/2. March ISLE MD workshop to discuss first IMDI proposal
  - Interested and committed people should come
  - Interested and committed people are invited to contribute

