From Grid and archive federation to CLARIN – Common Language Resources and Technology Infrastructure: networked technologies for linguists

Peter Wittenburg

MPI for Psycholinguistics

DOBES Archive

(Documentation of Endangered Languages)

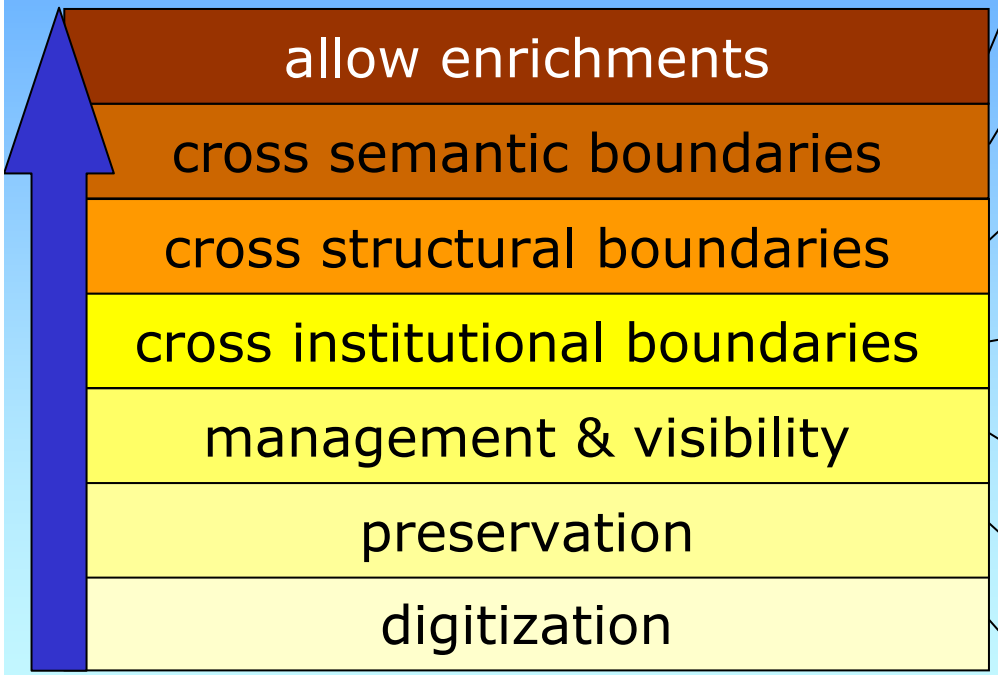
the major problems

- cultural, language and biodiversity is highly endangered
  (from the 6000 languages every second week one becomes extinct)
- according to a UNESCO overview about 80% of our recordings about cultures and languages are endangered as well
- the amount of high quality digital resources about cultures and languages is exploding
- according to the speaker > 80% of our digital material on notebooks and PCs is highly endangered
- according to the speaker ?? % of our digital material is not organized, catalogued - therefore invisible and thus inaccessible
- researchers have to cope with all sorts of boundaries and restrictions *(commercial, institutional, structural, semantic, etc)*

the challenges

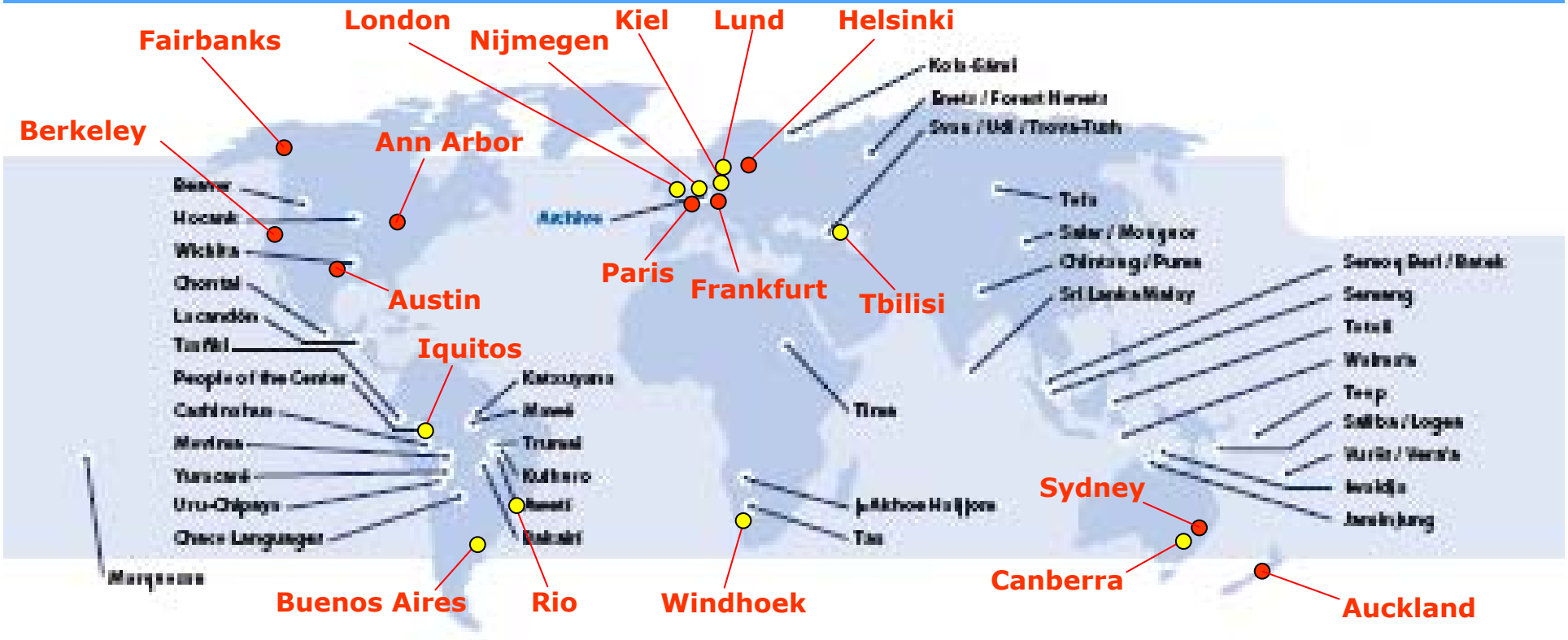
- maintain diversity (and revitalize where possible) – not issue today
- digitization programs and centers storing digital resources
- preserve cultural heritage for future generations
  - not simple, not a technological issue, but organizational + political + luck
  - technological solution not new: continuous migration + copying
  - cost need to be extremely low (10% law for bit-stream preservation)
- manage large amounts of data and make them visible
  - standardized XML schema based open metadata
- allow users to access data across institutional boundaries (Grid)
- overcome structural and semantic interoperability problems when accessing data (Semantic Web)
- allow enrichments according to the Live Archives ideas
- this all means support of the Open Access idea

layered system of dependencies



- it's the basis for knowledge preservation
- it's the basis for complete content disclosure
- it's the basis for content classifications/grouping
- it's the basis for virtually joined collections
- it's the basis for access
- it's the basis for continuity
- it's the basis for all work

example: documentation of End. Lang.

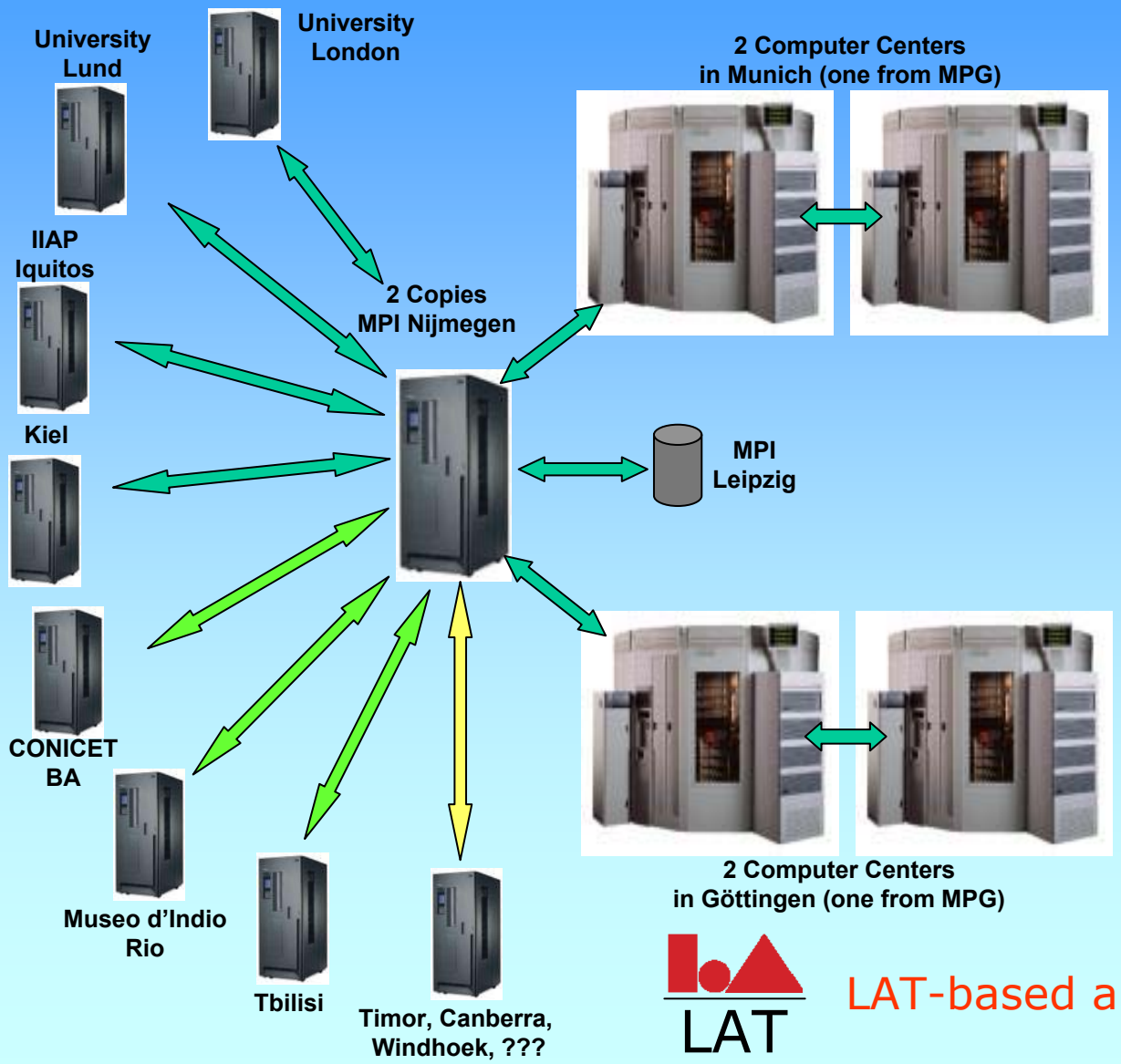


- DOBES Programme of VolkswagenFoundation (36 teams)
- other such centers for digital archives exist/emerge (DELAMAN)
- all are collaborating – some in a Grid (DAM-LR+)

GIS as the meeting point

- GIS “Language Sites”as an excellent meeting point

long-term preservation + distribution



- at MPI about 25 Terabyte = 25.000 h media
- > 250.000 objects
- 50 Mio annotations
- all types of linguistic data types (primary recordings, annotations, lexica, ...)

- fully described by IMDI metadata descriptions
- major types in open formats (XML schemas, lin PCM, MPEGx, ...)

LAT

LAT-based archives

Language Archiving Technology

LAT

Shoebox

CHAT

Transcriber

some XML?

ELAN / LEXUS / SYNPATHY
Media Annotation + Lexicon

IMDI
Data Organization
Metadata Description

LAMUS
Data Uploading and Management
Access Management

Data Archiving and Copying

IMDI
Metadata Browsing & Searching

ANNEX/LEXUS/ADDIT
Advanced Content Access via the
Web incl. content searching

offline loop for
manipulations
in the field

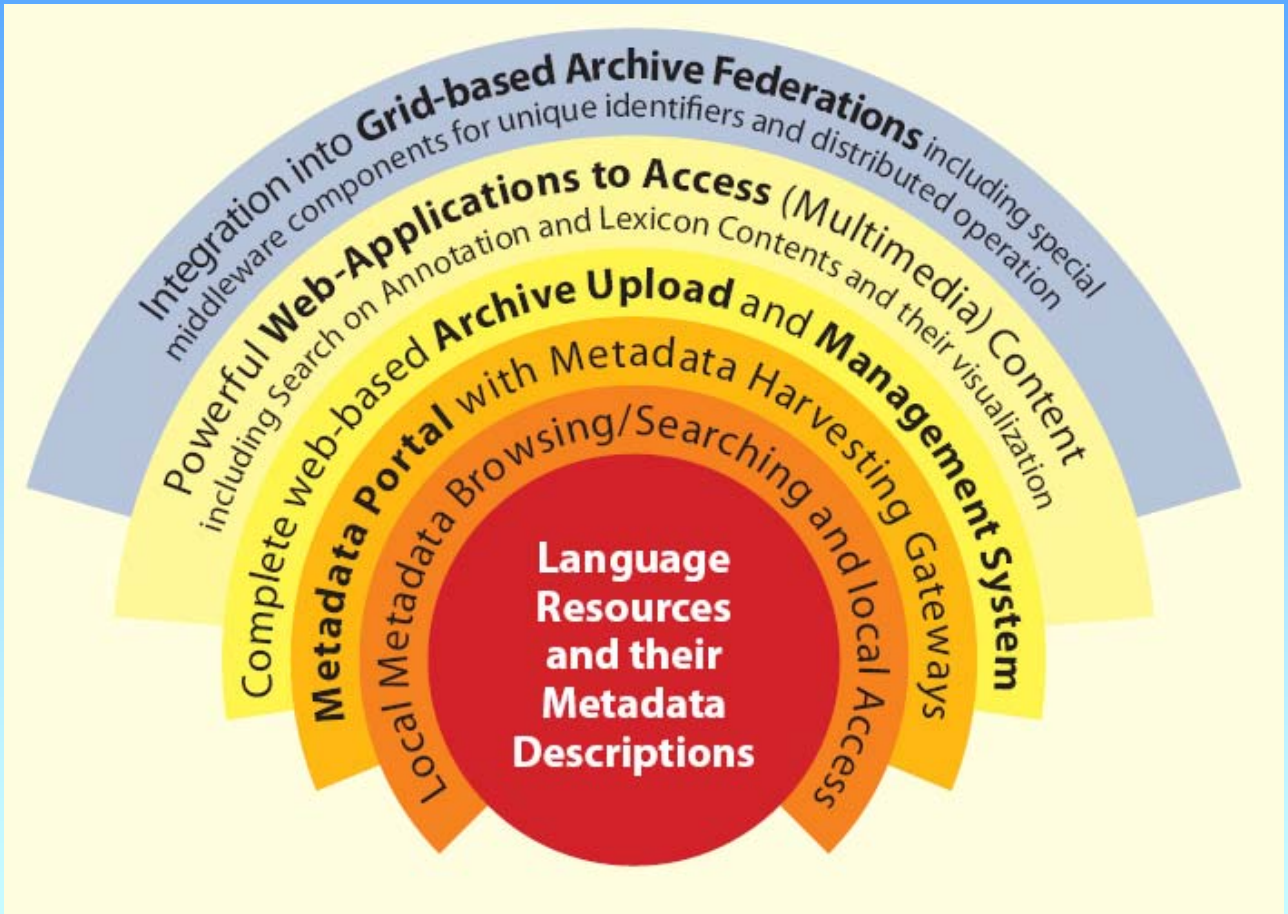
online loop for
manipulations
via the web

GRID
Extension

ODIT
Ontology Editor

Knowledge
Ontology
management
framework

from field to centers



you can start on a notebook

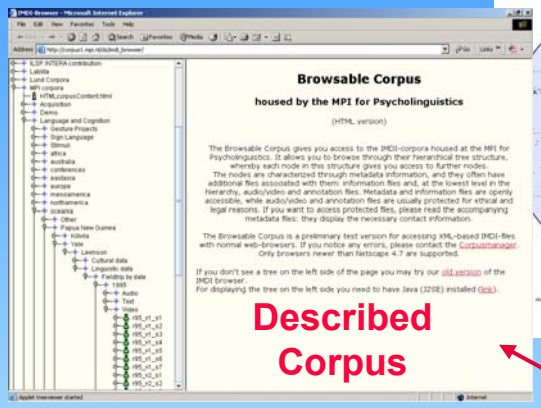
create your own small repository

and then upload it to a server

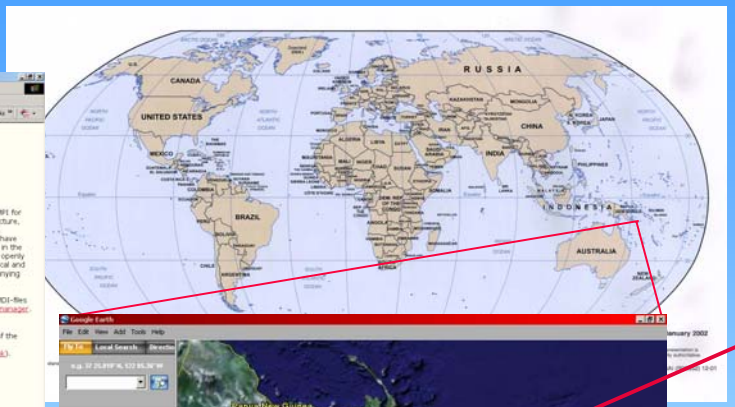
all is immediately accessible via the web

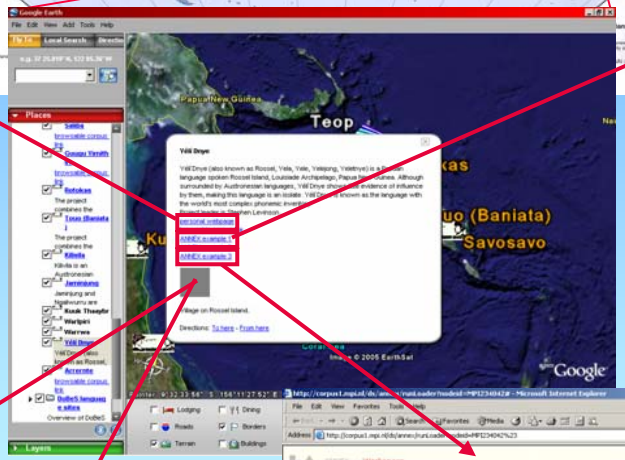
broad range of web-based access tools

LAT



Described Corpus



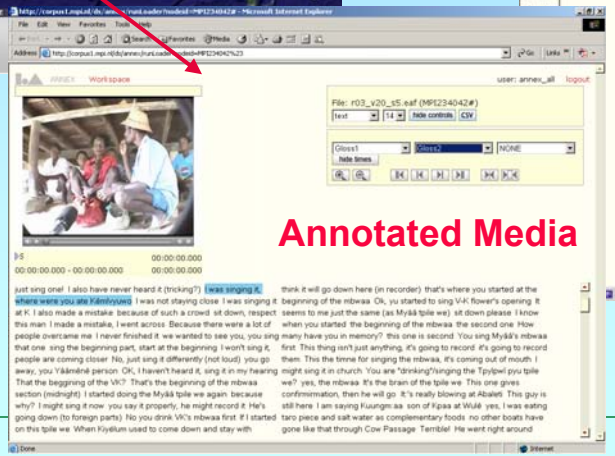


Multimedia Lexicon

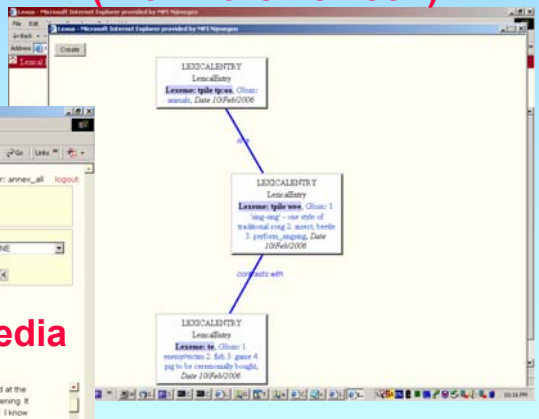
Typed Relations (within the Lexicon)

Video Clips

Photos



Annotated Media



crossing institutional boundaries

Goals and Federation

DAM-LR is integrating the language resource archives (LRAs) of the partner institutions so that they appear to users as one single large repository. The DAM-LR partners support the „Live Archives" principles for Digital Language Resource Archives.

Goals

The goals of the DAM-LR project are to create an integrated and unified domain of:

- trusted servers and services
- deep metadata for research purposes
- stable and unique resource identifiers
- user management and authentication
- exchange of user credentials for access authorization
- longer-term potential for exchanging resources to strengthen preservation purposes

Federation Platform

DAM-LR's integrated and unified domain can be called a Federation of LRAs, if the partners agree on issues including:

- a shared mission to provide integrated services
- mutual trust that the partners follow mutually agreed rules
- a common ethical and legal ground for all joint activities
- a number of practical guidelines such as which user credentials are to be exchanged
- the originating institute for a resource retains control of rights and access to it
- each LRA retains independence of operation

DAM-LR will establish a formal federating agreement in 2006.

DAM-LR

A 6th Framework Program Project DG Research

Partner Institutions

The DAM-LR project was started as a small scale project under the 6th Framework Program of the European Commission (DG Research) to introduce Grid technology and to virtually connect the language resource archives that are housed by the partner institutions listed below.

MPI for Psycholinguistics, Nijmegen (coordinator)

The MPI stores several types of linguistic resources (corpora, lexica etc.) collected by a variety of projects ranging from child acquisition, second learner acquisition, national spoken corpora, and sign language to endangered languages (DOBES). Currently, the archive holds more than 150.000 objects occupying over 15 terabytes. It is organized using the open IMDI metadata infrastructure.

Centre for Languages and Literature, University of Lund

The Lund centre houses a broad range of language departments and a high-end laboratory tailored to the study of cognition and language behaviour online. The centre relies on IMDI for the organization of its various language corpora - from child language development, dialects, field research in South East Asia - as well as its rapidly growing body of data from eye-tracking research (mostly reading), finger-tracking research tailored to the study of tactile reading, motion-tracking research (language and gestures), phonetic investigations, studies of online writing, and electrophysiological measures of reading, writing, speaking and listening activities.

SOAS, University of London

ELAR, The Endangered Languages Archive at SOAS, is a digital archive for worldwide endangered languages resources. The collection consists of linguistic and multimedia documentation materials deposited by funded researchers and others.

Institute for Dutch Lexicology, Leiden

The INL collects Dutch words. These words are stored in a database: The Language Database. The INL's TSTcentrale (language and speech technology centre) maintains and distributes state-of-the-art digital (Dutch) language resources such as the Dutch Spoken Corpus (CGN).

www.mpi.nl/dam-lr

DAM-LR

Distributed Access Management for Language Resources

A Grid Project

May 2006

- DAM-LR just established a Grid of LR archives
- it's about virtually merging collections:
  joint metadata, joint URID resolving, distributed access
- towards pan-European federation and beyond
- all LAT-based archives are easy to integrate

crossing structural + semantic boundaries

- some groups are working on generic models for LR types
- ISO TC37/SC4 working on generic models
  - Lexical Markup Framework with XML instance (workable)
  - Linguistic Annotation Framework
  - Morphosyntactic Annotation Framework
  - set of tools supporting multiple formats/schemas (LEXUS)
  - standardized Web APIs
- ISO TC37/SC4 working on semantic framework
  - Data Category Registry as concept reference framework
  - will not be sufficient
  - extension to multilayered compatible framework
  - at the bottom layer a flexible ontology editor
  - long way to go – need to start

towards common infrastructure

- what about flexible access of language technology and getting all the bits together?
- need to work on service oriented architecture
- implementing all layers requires large investments
- researchers will not rely on outcomes of short term projects
- for quantum leaps persistent research infrastructures necessary

some web-sites and end

Live-Archives:	http://www.mpi.nl/dam-lr/lra-flyer/lra.html
Open Access:	http://oa.mpg.de/
DOBES:	http://www.mpi.nl/dobes
MPI/DOBES-Archive:	http://corpus1.mpi.nl/ds/imdi_browser
Language-Sites:	http://www.language-sites.org
LAT-Technology:	http://www.mpi.nl/lat
DAM-LR:	http://www.mpi.nl/dam-lr
ISO TC37/SC4:	http://www.tc37sc4.org/
LIRICS:	http://lirics.loria.fr
CLARIN:	http://www.clarin.eu

Thanks for your attention