



Language Archives – essential pillars for eHumanities

Peter Wittenburg

MPI for Psycholinguistics



Research is in the Focus

- all we are doing **MUST** have intention to stimulate research!
- **research is in the focus!**
- however different approaches wrt interaction and cross-fertilization
- creating new stable infrastructures costs at least 5 years
- therefore at MPI all the years a pro-active attitude
 - may not look just to the direct needs researchers think of
 - have to anticipate what technology will offer in future
 - if no cross-fertilization humanities are always behind
 - more simple for hard sciences - since closer to technology
 - many examples at the MPI
 - started about 5 years ago with real archiving – now users start seeing the benefits and start relying on new methods
- **balance between short-term and mid/long-term goals is essential to meet the needs of the field**



Challenges in Linguistics (just a few)

- understanding the human mind and the way it processes language
- understanding how children and adults learn languages
- understanding, maintaining and documenting linguistic diversity
- improving our formal models and estimating their parameters based on ever growing and richer resources
- automatically analyzing media streams in a world where kids turn away from texts
- automatically translating texts and speech
- etc



Challenges in the Humanities (correct?)

- managing the huge problems resulting from immigration, i.e. all consequences of cultural diversity in our multi-cultural societies (NL, D, etc but also Europe)
- managing the societal balance given the extreme dynamics
- making knowledge available and transparent to more and more people as an essential basis of democracy
- move from a domain of hyperlinks to a web of knowledge, i.e. establish typed relations as a result of knowledge weaving mechanisms
- etc
- again: much more to say – just to convince ourselves



Competitive Situation

- obviously humanities proposals have to compete
 - fighting against cancer and other diseases (life sciences)
 - solving the energy problems of the future (hard sciences)
 - understanding how universe evolves (hard sciences)
 - etc etc
- well rating is about politics and one has to be careful with words
- but my personal view:
 - if we ignore the minds we will end up in chaos
 - live in fragile situation
- so yes we have to go for some money again and again and take actions to advance the humanities
- overcome fragmentation, crossing boundaries, ... -> e Humanities



What to be done I

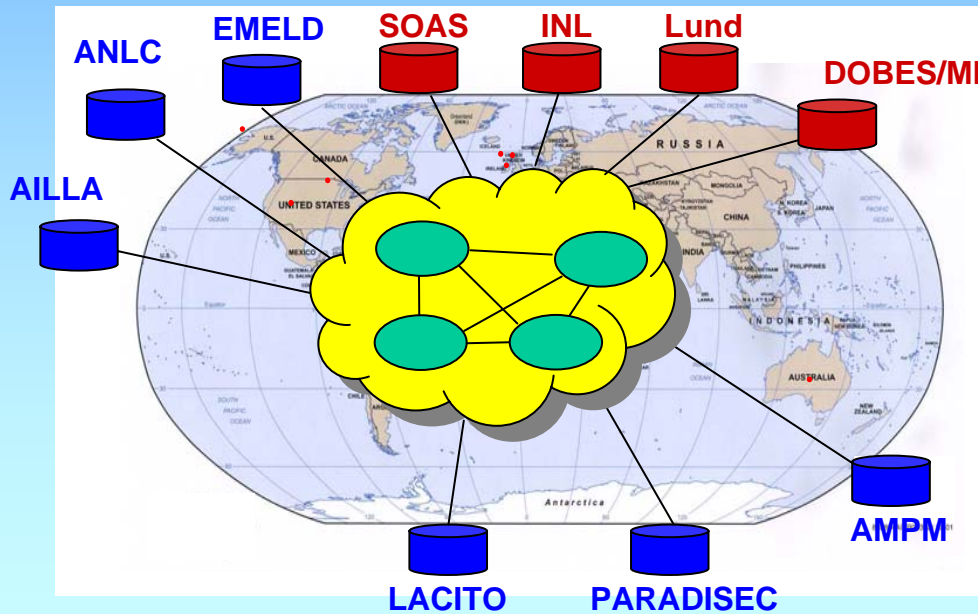
- obviously need new models, algorithms, applications etc (not today)
- obviously need to overcome current infrastructural limitations
 - **resource visibility** -> IMDI in language resource domain humanities often too critical

The screenshot displays two browser windows. The left window is the 'IMDI-BC Browser for MPI on md3403070', showing a metadata tree for 'MPI corpora' with various sub-categories like 'Acquisition', 'Language and Cognition', and 'Natural'. The right window is 'IMDI Browser - Mozilla Firefox', displaying the 'Max Planck Institute for Psycholinguistics: Browsable Metadata Domain' page. The page content includes a description of the institute's linguistic data collection, a list of goals (preservation, organization, availability, integration), and a list of MPI corpora. The right sidebar of the Firefox window shows a detailed metadata record for 'Project Aweti', including location (South America, Brazil), actor (Sebastian), and languages (German, English, Latin).



What to be done II

- obviously need to overcome current limitations
 - overcome resource+service integration gap -> Grid solutions
 - EU funded DAM-LR project
 - don't believe hard sciences that we don't know about things



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
- long-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.

A 6th Framework Program Project 06 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, HJensen (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 for endangered languages (DOBES). Currently, the archive holds more than 100,000 objects occupying over 15 terabytes. It is organized using the open DMOZ 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 DMILL 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), (input-tracking) research tailored to the study of teacher reading, action-tracking research (language and gesture), 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 multimedial 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 ISCentral (language and speech technology centre) maintains and distributes state-of-the-art digital (Dutch) language resources such as the Dutch Spoken Corpus (DSC).

www.mpi.nl/dam-lr

Distributed Access Management for Language Resources

A Grid Project
May 2006



What to be done III

- obviously need to overcome current limitations
 - need technical encoding interoperability (Unicode, MPEG etc)
 - need structural/format interoperability
 - generic schemas -> LAF, LMF, ...
 - need semantic encoding interoperability (hard problem)
 - tags and values -> ISO TC37/SC4 Data Category Registry
 - concepts in prose texts -> general ontologies, wordnets, ...
- we are working on these issues, but long way to go
- don't believe hard sciences; we understand this much better



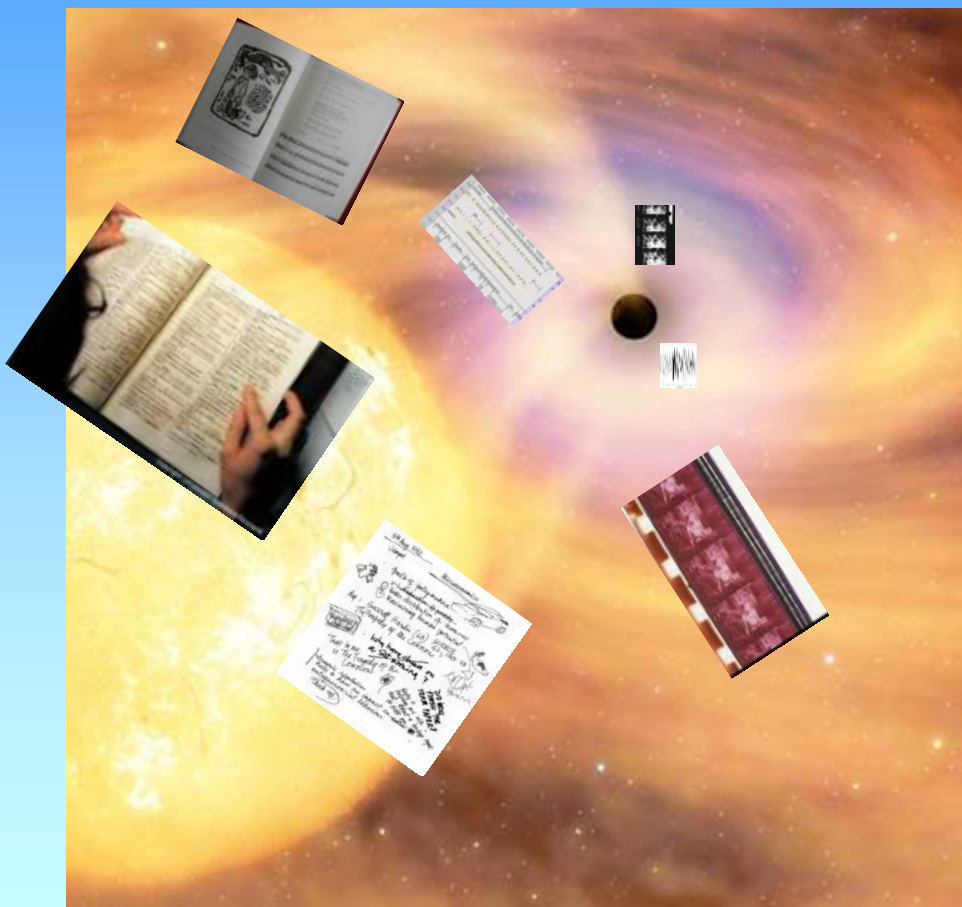


What to be done IV

- obviously need islands of stability and power to ensure
 - persistent accessibility of resources, services and knowledge
 - persistent availability of advanced services and registries
 - long-term preservation of resources
 - etc
- education+training, education+training, education+training,
education+training, education+training, education+training, education+training, education+training
- who can take care of this?
 - well call them centers or archives or ...
- but what are they?



The Researchers Nightmare



- the digital archive as a black hole
- valuable resources are “eaten”
- only contracts, procedures, license agreements are produced
- data seems to be inaccessible
- old business models to be found
 - limited user interfaces
 - just distribution of static resources
- etc etc
- **is this reality?**
- why are people getting more and more upset?



The Researchers Ignorance

- D. Schüller (UNESCO):

80 % of our recordings about languages and cultures are endangered!

- absolutely inappropriate treatment
- what about all the data encapsulated in relational databases?
 - they will die with the PC or software
- what with all the specially made CDROMs?
 - it's nice but useless for further research work
- what with all undocumented and unorganized data on notebooks?
- etc etc
- some speak about the danger of loosing our cultural memory



The Live Archives View

- be as open as possible (Open Access)
- open interactive&commentary services (greetings from Wiki)
- offer fragments from personalized collections dependent on the user
- the creators have free and unburocratic access
- researchers have flexible access
- is this reality? NO

Digital Language Resource Archives

User Services

Live Archives offer three core classes of user services: upload, management and access.

Upload

Digital archiving is an ongoing process. Depositors and users can upload additional resources or update existing ones. Data management systems will provide facilities for upload, while tracking versions and identifiers so that important information and references will not be lost.

Management

LRAs data management systems allow users to define access policies and rights, to carry out checks, to do conversions and downloads, and to view statistics about their data and the rate of access.

Access Services

LRAs will offer a number of access services:

- WWW catalogue browsing and searching via metadata or via geographic interfaces, to support resource discovery
- WWW access to resources (where permissions allow) via standard web browsers
- individual resources as well as sub-corpora, together with relevant metadata, should be downloadable to a variety of devices
- specialized applications to allow complex resources to be accessed via the WWW
- open-standard metadata (OAI-PMH, XML-Schema) provided for harvest by other bodies to expand resource discovery potential

Supporting Institutions

The intention is that as many institutions as possible that hold language resources will support the Live Archives principles and directions, as described in this leaflet. If your institution may be interested in supporting it, or finding out more information, please fill in the form on the web-site:

www.mpi.nl/dam-lr/lra-flyer

Currently, the following institutions have declared support for Live Archives (see the website for the latest information):

- MPI for Psycholinguistics, Nijmegen*
- Centre for Languages and Literature, University of Lund*
- TSI-INL, Leiden*
- SOAS, University of London*
- Radboud University Nijmegen University Sheffield
- Charles University Prague
- Bulgarian Academy of Sciences, Sofia
- Berlin-Brandenburgische Academy of Sciences
- University of Bergen
- Istituto di Linguistica Computazionale, CNR, Pisa
- University of Antwerp
- Centre de Ressources sur la Description de l'Oral, CNRS, Paris
- University of Helsinki
- Institute for Language and Speech Processing, Athens
- University of Zagreb

Other institutes that declared their support for this document are listed on the web-site.

Digital Language Resource Archives

LIVE ARCHIVES

A checklist of principles and tasks

March 2006

*The Live Archives statement was initiated by the DAM-LL partners (www.mpi.nl/dam-ll).

little PR:

support the Life Archive ideas !!

have to change the rules of the game

have to create awareness

comments welcome



Give it a try at the MPI

- short non-commercial overview
 - MPI has a language resource archive covering now
 - ~23 TB (annual increase about 8 TB)
 - ~250.000 resources
 - all described (with sometimes lousy) metadata
 - all organized and manageable
 - most resources in archivable formats (long-term)
 - all accessible via the web
 - a number of different discovery methods
 - a large variety of access methods
 - from bulk upload/download to fragment access
 - integrating some archives (EU funded Grid project)
 - are working on “ontology” support and commentary tools
- **is this reality?** - well are working hard ;-)



The Landscape I

- urgently need centers/archives as islands of stability
- specialization in services and knowhow seems to be necessary
- MPI knows something about approaches in LRT domain and the appropriate media types
- humanities are much much broader of course
- but all need/use LRT – so LRA have an important role
- helping tackling the Gand Challenges -> eHumanities scenario
- do we need some form of a glue and added value?
- do we need more than LRA?



The Landscape II

- do we need some form of a glue and added value?
- do we need more than LRA?
- my personal answer is yes
- but
 - keep research in the focus
 - remain sensitive to the researchers
 - but don't forget the mid/long-term issues (not always rated)
 - re-use existing knowhow
 - bring together instead of separate field specialists
 - build upon what is already there
- support the DANS model – but we have to make it concrete



The Wishes

- so my best wishes for the success of DANS and good luck for Peter and his crew
- will need it – researchers are often impatient and egoistic
- let's hope on a fruitful collaboration

eHumanities Workshop within eScience Conference Amsterdam

- *December 4-6*
- *paper submission until August 20*
- *www.mpi.nl/clarin -> events*

thanks for your attention