



Databases for Linguistic Purposes

Peter Wittenburg, Daan Broeder, Kees vd Veer
Max-Planck-Institute for Psycholinguistics
Richard Piepenbrock
Nijmegen University



Intention

- looking back at almost 20 years of database applications
- what did we do – why did we do it?
- does it make sense what we did?

- not an easy enterprise for data driven people
- sat together a few times to answer these questions

- 1. will briefly introduce some work we have done
- 2. will look out to what will come next
- 3. will share a few conclusions

- copied some schemas



What kind of databases do we have?

- administrative data to organize linguistic work (example)
 - equipment situation
 - journal administration
 - ...
- experimental data
 - time series, rt data
 - numerical, structured, constrained, simple metadata
 - sequential & statistical processing
 - special file formats
- observational data
 - av recordings
 - various channels (speech, gestures, eyes, ...)
- linguistic data (from various domains)
 - annotations
 - lexica
 - various notes (typically unstructured and mixed data)
 - metadata
 - ...



TGORG application

- Technical Group Organization Database
- running since 1985 – created in the early phases of rDBMS
- built on ORACLE as a typical relational DB application
- the core to administer all our equipment (control, planning, ...)
- a number of clear administrative entities such as equipment units, equipment types, users, hubs, ...
- shared by all responsible TG members
- planning of about 30 expeditions with different report types
- beginning goal was to be prepared for the CELEX project
- test bed to use all good ORACLE features (constraints, triggers, ...)
- turned out to be the oldest and most stable application at MPI
- funny: central control people first complained
later our solution was sold as an example to others



CELEX Application

- first big computer-based lexicon project in NL
- started in 1985
- goals
 - create computer-based lexica for D, G, E
 - offer interactive access to researchers
 - include all types of lexical information except semantics
 - so also frequency counts generated on large corpora
 - change the way of creating lexica and working with them
- working on computers meant to create a formal model
- after intensive analysis work and discussions decided to use the relational model as basis
- received much critique from linguists
 - relational model too simple to represent linguistic complexity
 - have seen the shelves at INL full of cards with notes understand partly what is meant



CELEX Application

- together with CS from TU Eindhoven development of LS
- after some discussion rounds and adaptations the LS was accepted as “holy core”
- work could focus on ingestion, merging, correction, ...
- much programming around SQL core
- needed many procedural components – embedded SQL
- for D about 40 tables and 400.000 full forms
- access via alpha-num terminals (semi-graphical)
- users could create temporal private tables
- one of the most frequently used tools in linguistics in NL etc
- problems:
 - some calculations took much CPU time (neighbors, uniqueness, ...)
 - storage space was limited
- later: some people wanted to work self-supporting etc
 - created a CDROM with simplified tab-delimited tables + Perl scripts
 - have a simple web-site without support

www.mpi.nl/world/celex



Speech-Error Database

- 2002 we received a request to create a unified SE DB
- speech error registration is a kind of hobby of some researchers
- they listen, hear something funny and write it down on paper
- all in individualistic styles and often with little information
- some of this exists on computers
- useful to study speech production and self monitoring processes
- in general:
 - error as orthographic string, sometimes phonetic
 - target with several options (ambiguous)
 - language and date
- intention: unify different SE DB and make it web-accessible
- procedure:
 - linguistic analysis of attributes
 - mapping were possible
 - design of an exhaustive XML schema to not loose data
 - with scripts creation of one XML file (now 8600 entries)



Speech-Error Database

- Question for us:
 - how to make it web-accessible?
 - searching should be fast
 - did not want to invest too much time
 - tested XML DBMS (eXIST, ORACLE 9i, ...) at that time
 - results were frustrating (bugs, little speed up)
 - decision to transfer XML file to relational DB (Postgres)
- Problem:
 - structured data but sparse filling and many 1:N relations
 - object-relational mapping would lead to many small tables
 - only some major attributes were selected to be searchable
 - joins just for data presentation would slow down search
 - therefore, many attributes as one XML/HTML structure
- so in total not a nice solution – against all recommendations
- it's available on the web with simple UI

www.mpi.nl/corpus/sedb

(unofficial)



Metadata Database

- the IMDI domain is a distributed domain of linked XML files adhering to the IMDI Schema
- used for management and discovery purposes
- MD files are at different centers (MPI, Lund, BAS, ...) and on PCs and Notebooks (fieldworkers)
- is it a database – yes, but ...
- simply connect to the web and register the node
- it is an open well-documented domain
- distributed domain is visible with IMDI Browser (HTML to come)
- if you know the URL you can access all MD (create own services)
www.mpi.nl/corpora
- OAI model is different:
 - any repository can have its own MD set
 - providers deliver data according to a schema (DC, OLAC, (IMDI), ...)
 - result is a searchable index



Metadata Database

- why did we do so
 - low threshold for everyone who likes IMDI (not per se)
 - all in archivable format and part of the archive
 - no encapsulation, i.e. direct access
- Problems?
 - browsing not a problem (IMDI browser, XSLT trafo to HTML)
<http://corpus1.mpi.nl/BC/IMDI-corpora/>
 - searching requires harvesting and indexing
 - currently > 30.000 MD descriptions of linguistic units at MPI
 - ~ 100.000 objects due to bundling for mm recordings (~ 8 TB)
 - further 20.000 MD descriptions ready from other sites
 - first solution (text index + Perl scripts) did not scale beyond 10.000 MD
 - now use of Java rDB library – is ok so far
 - why not ORACLE or POSTGRES?
 - for local work an installation and requirements problem
- in ECHO (>150.000 MD) tests with binary tree index
corpus1.mpi.nl/ds/dora



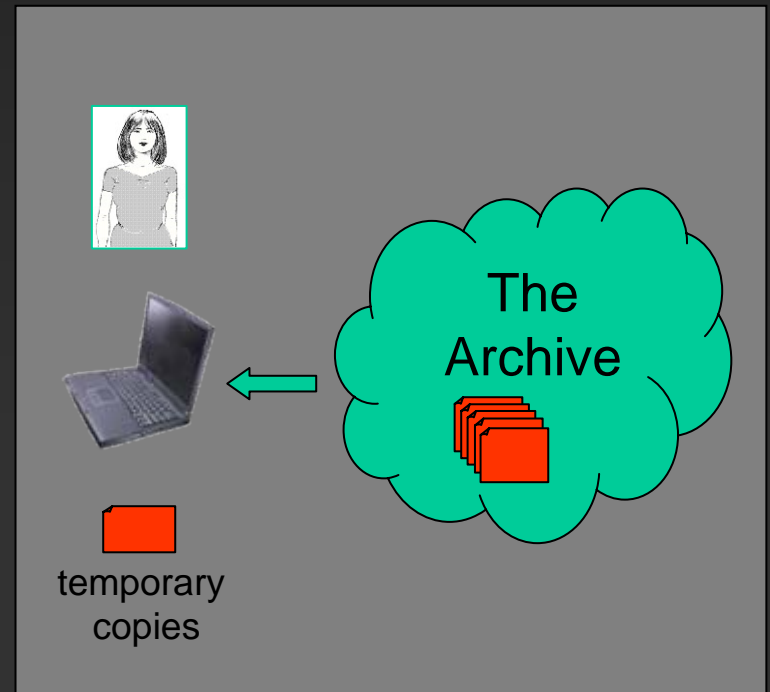
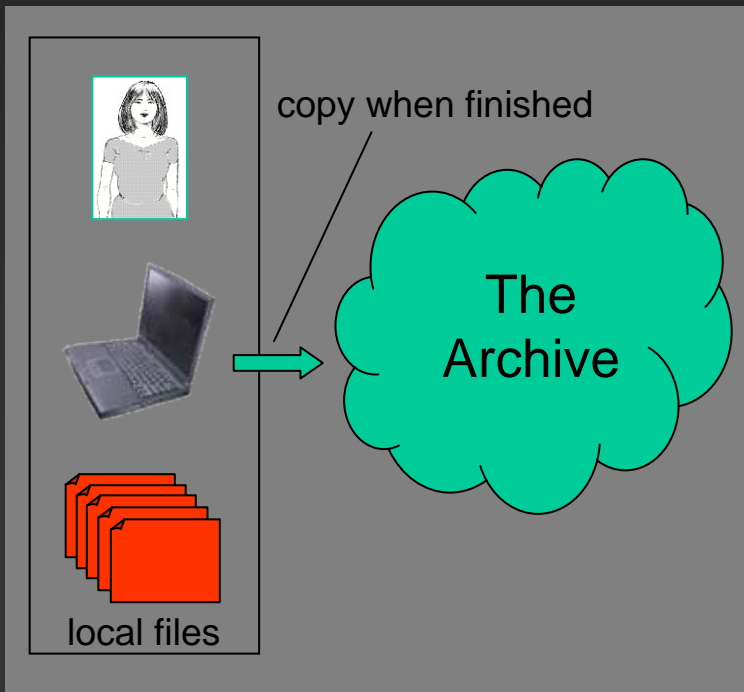
Metadata Database

- is the solution ok?
- distributed XML scenario and long-term archiving is primary focus
- searching and speed is secondary focus (derived data)
- Pros:
 - no data encapsulation – archivable format
 - no platform dependency
 - no special DBMS needed
 - naturally distributed
 - domain integration and openness very simple
- Cons:
 - need IMDI Browser or XSLT trafo to work on domain
 - need harvesting for searching



MPI Archive

- as indicated: various linguistic data types in the archive
- variety of different types of relations amongst the objects
- some (at object level) can be modeled by IMDI metadata
- the archive is accepted as something comparable to an accelerator engine in physics – the core research instrument
- the perception of our researchers changes



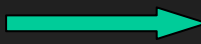


MPI Archive

- is the archive a database? yes but ...
- our choices as mentioned before:
 - no encapsulation for archival objects – direct accessibility
 - all in readable formats where possible (XML, plain text, ...)
 - all uncompressed where possible (video not yet)
 - all archivable
 - all part of the same copying mechanisms
- now we need better access and exploitation tools



Archive Exploitation

- Current Questions:
 - How to extend ELAN search on several/many EAF files?
 - How to do search on The Archive?
 - How to flexibly visualize and combine objects from the archive?
first attempts made 
will get a grant together with MPI Leipzig
- ELAN allows to create and exploit mm annotations
 - currently complex search on one EAF/XML file!
 - ELAN is a local tool!
 - Multiple-file search with ELAN requires index as well
 - so same question: what to do on a local machine?
- Archive search is a central component, i.e. no problem to use rDBMS for fast searching (ORACLE not acceptable)
- but what with unstructured documents?



Web-based exploitation

Browsable Corpus - Microsoft Internet Explorer

Address: <http://corpus1.mpi.nl/BC/IMDI-corpora/MPI%20corpora/Demo/PeWi%20corpus/kleve-route?annotation>

(HTML version)

History: [IMDI-corpora](#) / [MPI corpora](#) / [Demo](#) / [PeWi corpus](#) /

Current point in the tree:

Information available at current point of tree:

This recording was made to generate a freely available test resource including speech and gestures. The annotations were created by Peter and Kita who is gesture researcher at the MPI for Psycholinguistics.

Diese Aufnahme wurde erzeugt, um eine frei verfügbare Test Resource zur Verfügung stellen zu können, die Sprache und Gestik umfasst. Die Annotationen wurden von Peter und Kita, dem Gestik Researcher am MPI erzeugt.

- Metadata
- Media
- Annotation

[elan-example1_eaf](#) (download eaf-file)

[elan-example2.smil](#) (0.0 MB)

[elan-example4.smil](#) (0.0 MB)

Class EAF HTML Viewer

Annotation	Start Time	End Time
in from here	0:00:00	0:01:40
yeah	0:03:20	0:04:00
in	0:09:30	0:10:00
in	0:14:30	0:15:00
there is another path	0:22:30	0:23:30
rotunde	0:27:30	0:27:30
do you go out of the bookstore to the Saint Anna Street	0:28:40	0:29:30
and then you take other, Saint Anna Street to get to the center of the town, to the big rotunde	0:40:00	0:10:70
and you follow them the sign-leaf	0:13:20	0:13:30
that's the orange single	0:13:30	0:14:40
then you follow the sign-leaf	0:15:30	0:17:50
yeah	0:17:50	0:18:00

RealPlayer

K-Spch

W-Spch

and then you go the other, Saint Anna Straat to this to the center of the town, to this big rotunde.

(Paused) elan-example2 (K-Spc) 0:09 / 0:36

Real Guide Music & My Library Music Store



SMIL style
media + subtitles



Archive Exploitation

- Current Questions:
 - How to extend ELAN search on several/many EAF files?
 - How to do search on The Archive?
 - How to flexibly visualize and combine objects from the archive?
first steps made
will get a grant together with MPI Leipzig (FIELD an option?)
- ELAN allows to create and exploit mm annotations
 - currently complex search on one EAF/XML file!
 - ELAN is a local tool!
 - Multiple-file search with ELAN requires index as well
 - so same question: what to do on a local machine?
- Archive search is a central component, i.e. no problem to use rDBMS for fast searching (ORACLE not acceptable)
- but what with unstructured documents?



What is coming next?

- just a few wishes
 - improve the synchronization between local and central copies
 - integrate archives
 - integrate user domains } DELAMAN, DAM-LR
 - increase semantic interoperability (DCR, Ontologies, ...)
 - create relations and exploit them
 - allow collaborative annotation and commentary (panel at LREC)
(have an ELAN prototype for collaborative video annotation
will be on the web for tests and comments)
 - assure long-term persistence (now 5(7) copies of relevant data)
- impressive list – how can we manage to create stable and robust systems?
 - did not yet achieve interoperability at encoding and structure level (ECHO example)



What is coming next?

- continuous stream of new technologies and solutions
- not at all clear where to rely on – we are part of the “evolution machinery”
- just a few technologies
 - abstract models such as LMF (ISO)
 - various container types (SRB, CMS, ...)
 - lot of data mining solutions
 - RDF/(S)/OWL simple relational model and framework for formalizing semantics
 - **web-services** to increase interoperability
 - stack of specifications (SOAP, WSDL, UDDI, Policies)
 - Open **GRID** Service Architecture/Infrastructure
 - **GRID Middleware** components
 - distributed URID services
 - distributed user/group management
 - security services (certification, authentication, ...)
 - **new Client SW** (Flash, SMIL, ...)
 - ???



Conclusions

- discussed pros and cons of XML/rDBMS
 - for us archiving requirements are primary
 - DBMS for special purposes at least for central services
 - distributed scenario important for us (WS change the game)
- may not underestimate the “real” problems (Gary’s points)
 - 80% of all av recordings about heritage are stored on shelves like books (Schüller)
 - how can we take care that a fraction will survive?
 - linguists create lots of excellent stuff using rDBMS, ... on their PC
 - how can we take care that a fraction of it will survive?
 - how can we come to a coherent archive?
- don’t know whether we made it right – miss useful criteria
 - short term wishes vs long-term needs
- things become comparatively simple if project-approach is taken



Something remaining?

message to Helen/Tony:

(almost) no best practice advice 😊