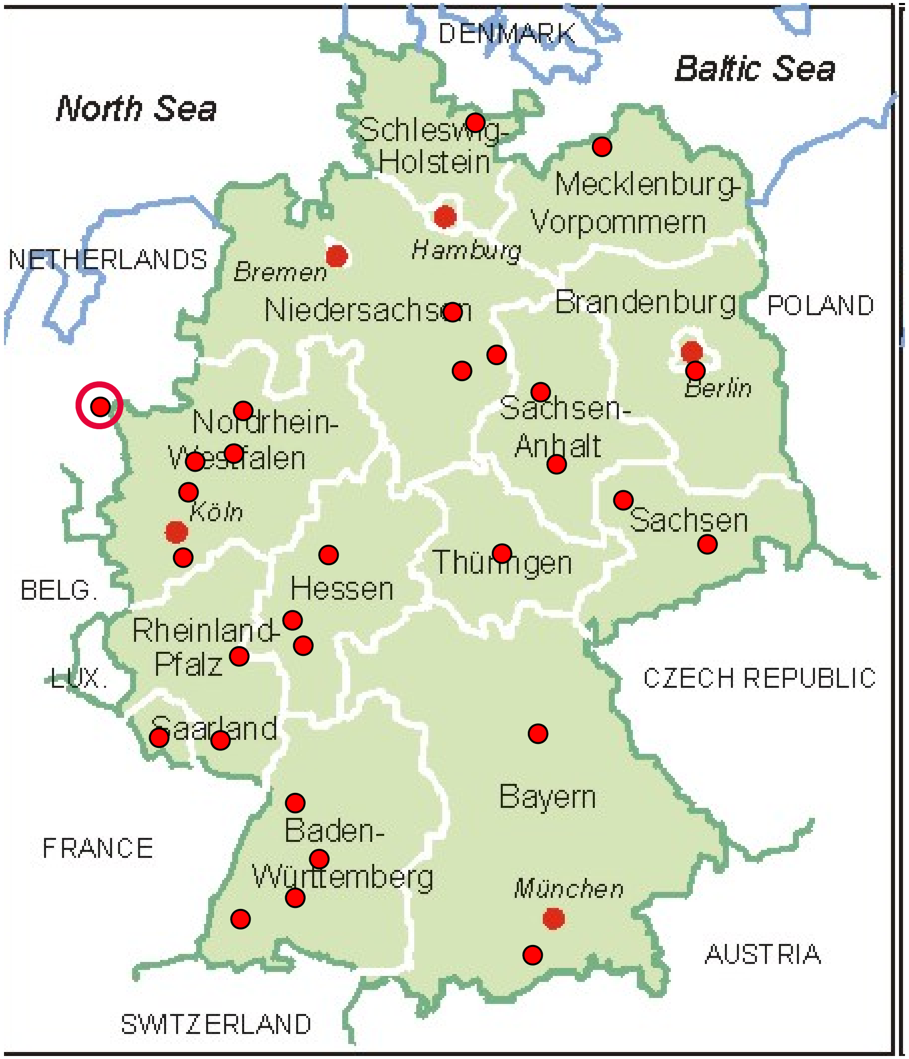
Culture Change in Data Management

Peter Wittenburg
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
CLARIN - European Research Infrastructure

where will I talk about?

- who is he - some background
- data management - MPI activities
- some related CLARIN activities
- major data management dimensions
- some basic IT principles
- summary

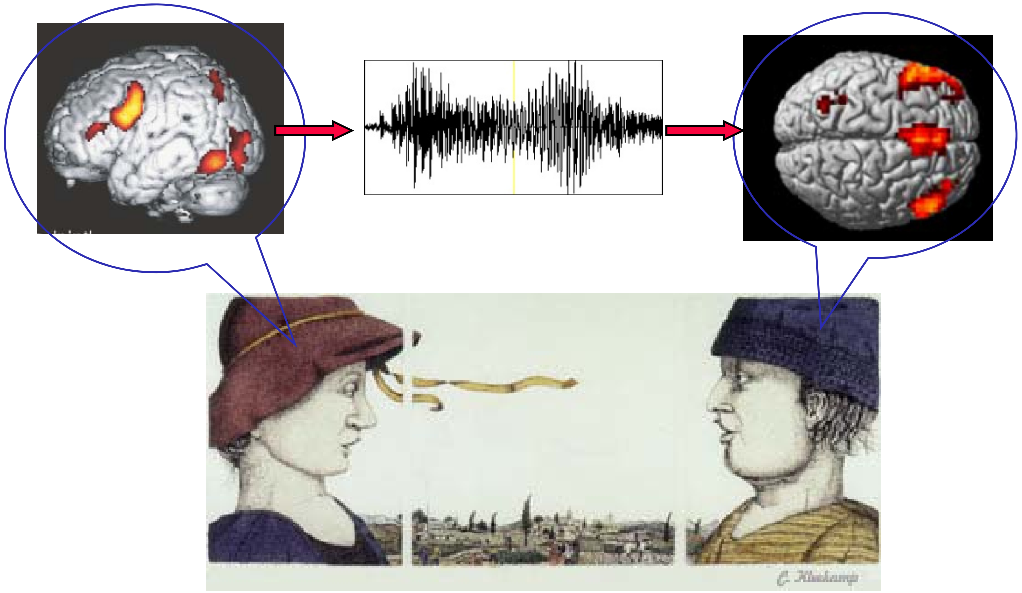
personal background - Max Planck Society



- > 80 Institutes for fundamental research
- mainly in Germany, a few in other countries (NL, It, etc)
- covering all disciplines
  - natural sciences
  - life sciences
  - social sciences
  - humanities
  - law
- personally: technical director at one MPI, member of the MPS IT advisory board

personal background - the institute

- languages: unique experiments of nature
- human mind: unique "creation" of nature to process NL
- at MPI fundamental research in mental language processing, language acquisition, language & cognition, neurocognition
- methods: experiments (VR), signal processing (eye movements, gestures), computer-simulations, brain imaging (EEG, fMRI), multimedia based observations of multimodal interaction



The Data Management Problem

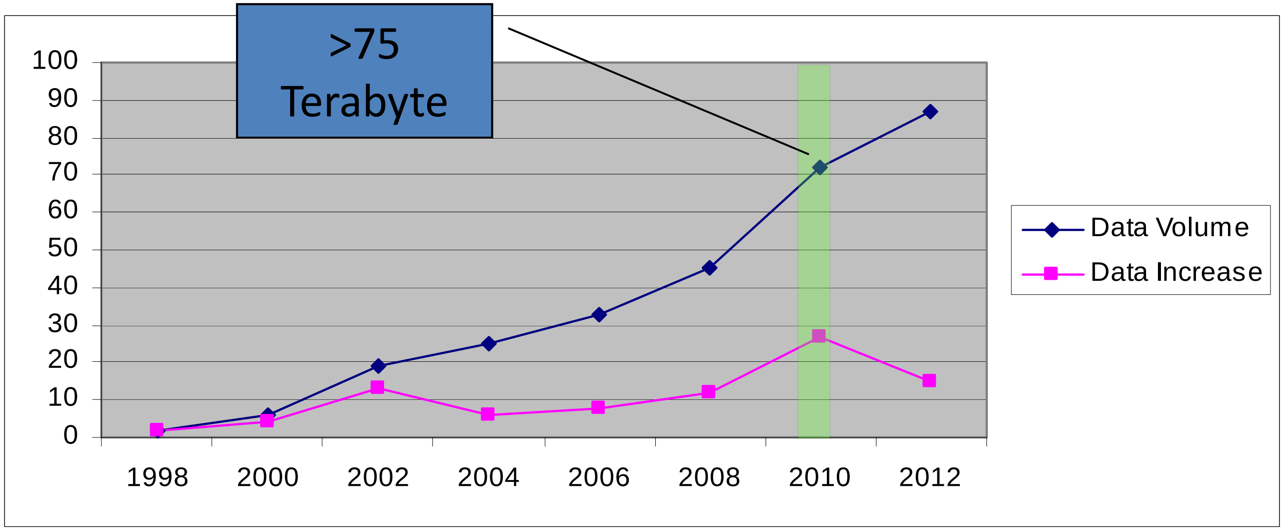
Data Management is the topic of my talk

first reality at MPI and related activities in CLARIN

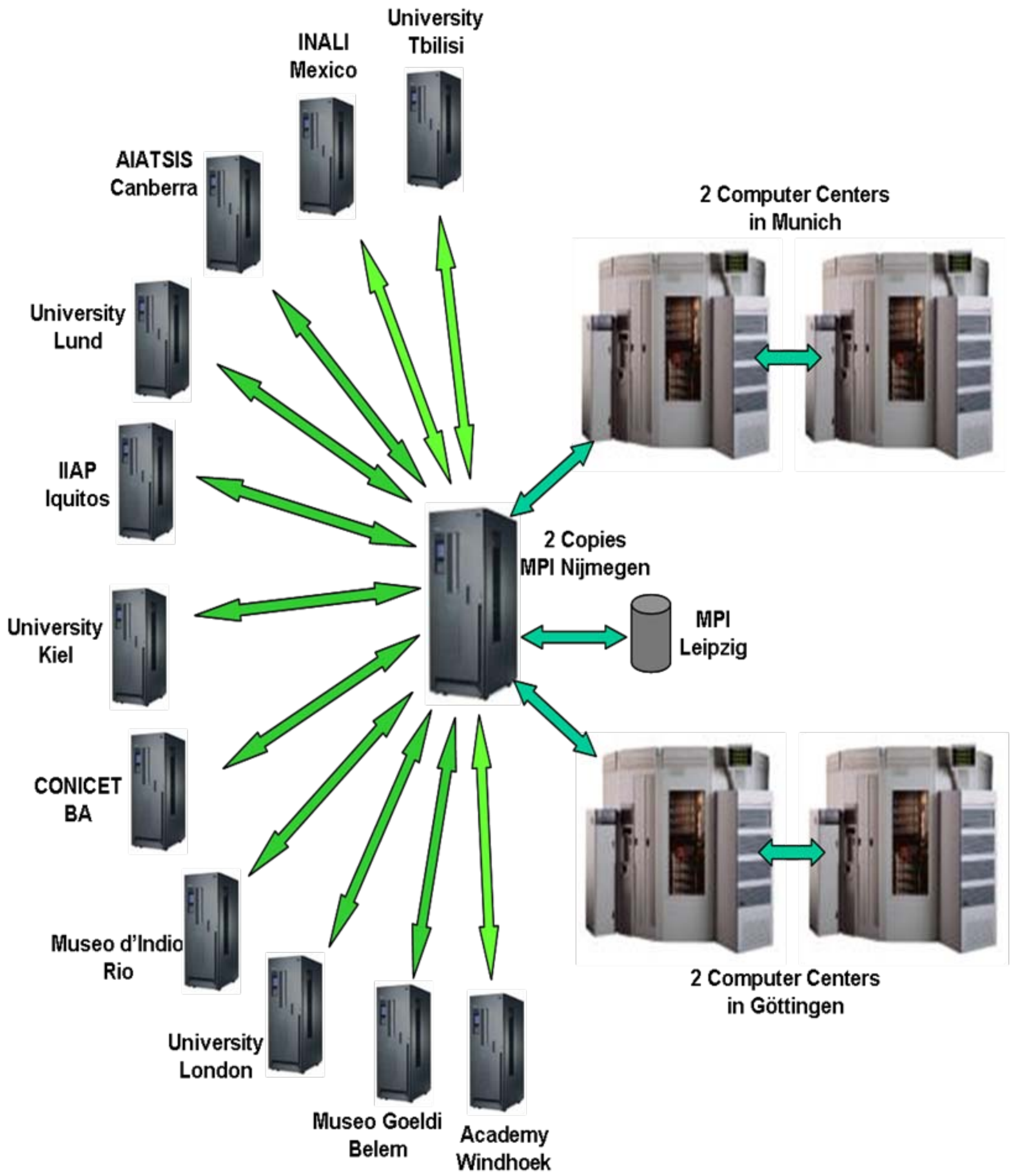
is it appropriate?
what are we doing?

data development at MPI

- MPI research is based on data - experimental and observations
- are we special - not really except perhaps
    - all-digital-world very early incl. digital audio/video
    - had to cope with lots of data already in the 90-ties
- two worlds: organized archive vs. huge MPI data backyard

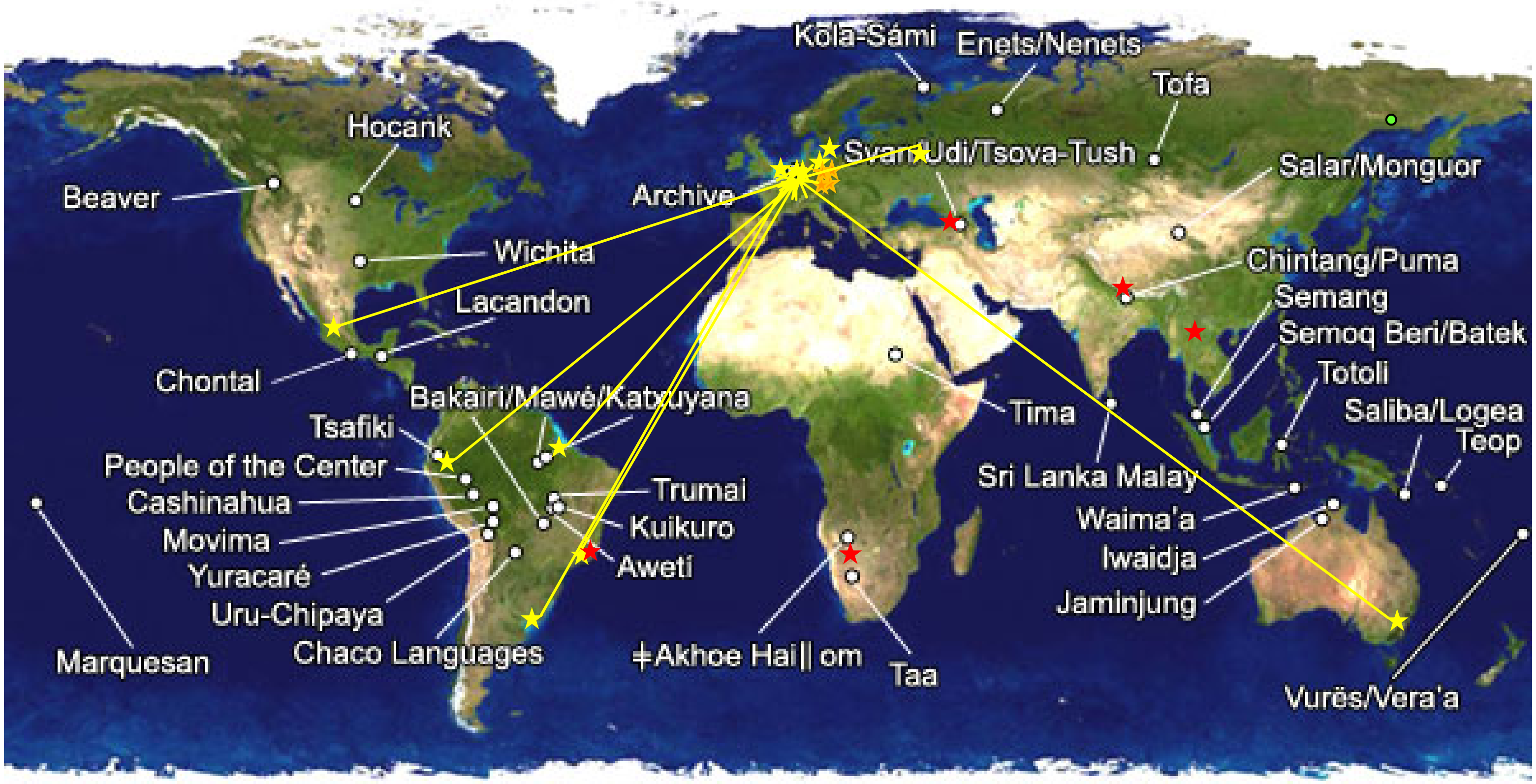


The Archive (DSA quality certificate)



- in total more than 200 TB of data - can store about 1 PB
- stable, robust, organized and coherent online archive with 50 Terabyte of resources ready for eScience
- all metadata described and all associated with PIDs
- 4 full dynamic copies at remote CC with 50 years guarantee
- in addition 11 regional repositories with more to come
- open deposit service

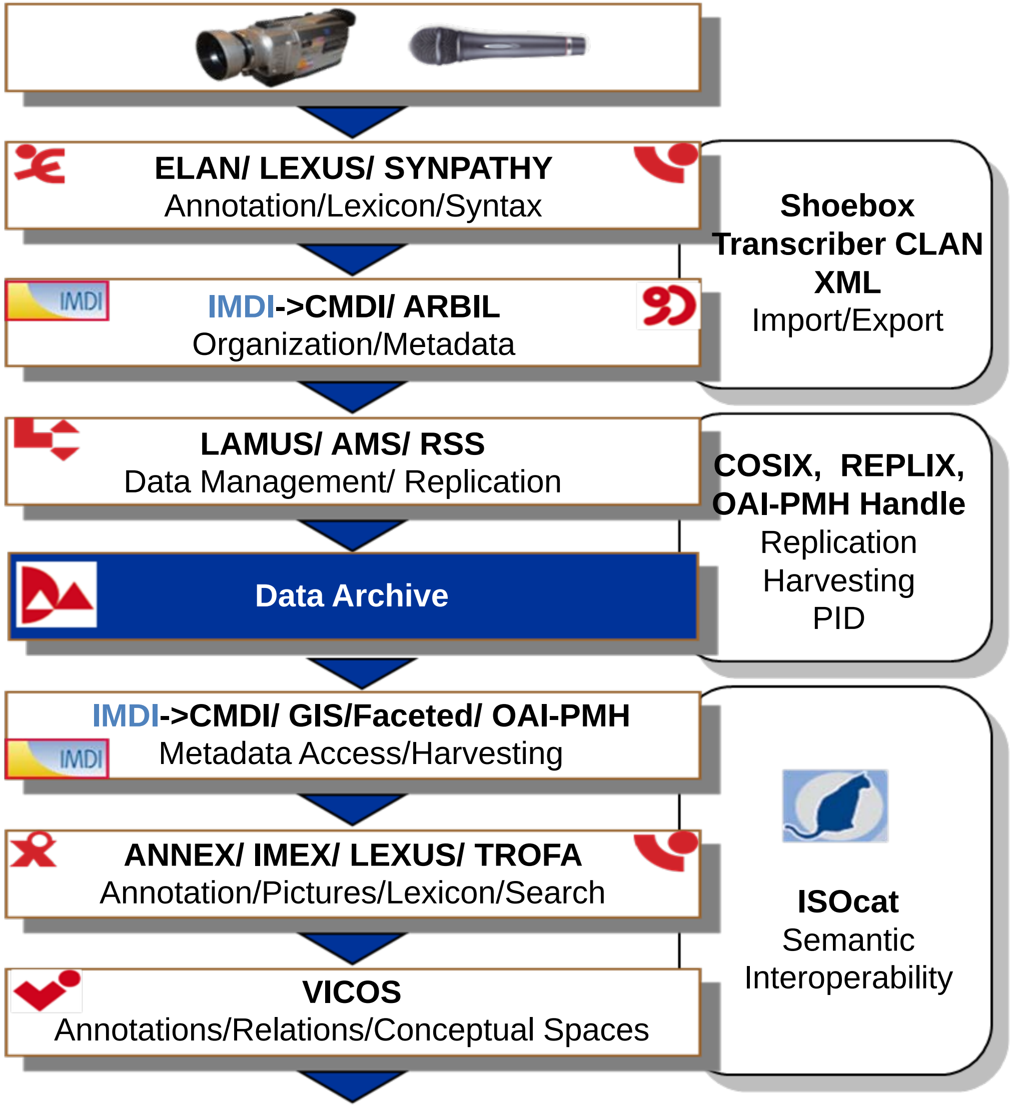
DOBES Project



- 46 teams all over the world documenting about 70 endangered languages
- currently every week one of the 6500 languages is dying
- unique treasure about our heritage to be preserved for future generations
- all annotations and analysis is manual - no NLP working

LAT Software Suite

The Language Archive



ELAN/ LEXUS/ SYNPATHY
Annotation/Lexicon/Syntax

IMDI->CMDI/ ARBIL
Organization/Metadata

LAMUS/ AMS/ RSS
Data Management/ Replication

Data Archive

IMDI->CMDI/ GIS/Faceted/ OAI-PMH
Metadata Access/Harvesting

ANNEX/ IMEX/ LEXUS/ TROFA
Annotation/Pictures/Lexicon/Search

VICOS
Annotations/Relations/Conceptual Spaces

Shoebox Transcriber CLAN XML
Import/Export

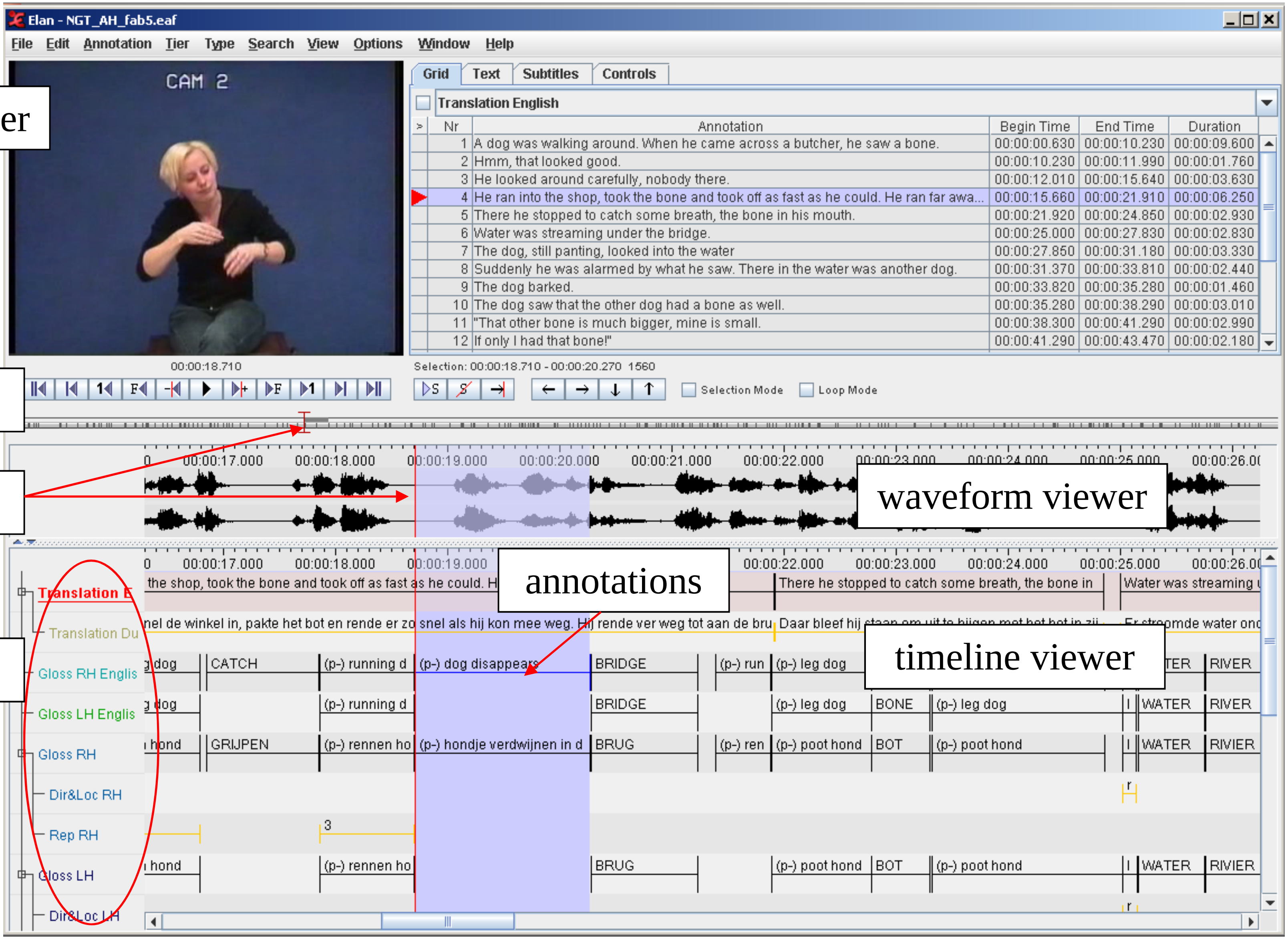
COSIX, REPLIX, OAI-PMH Handle
Replication
Harvesting
PID

ISOcat
Semantic
Interoperability

- full Lifecycle Support from data creation to semantic web like "exploitation"
- standards-based where possible
- modular design - all Java
- ELAN for example one of the most widely used annotation tools in the world
- data grid extensions
- ISO based interoperability extensions

professional annotation tool

The Language Archive



Elan - NGT_AH_fab5.eaf

☐ Translation English

Nr	Annotation	Begin Time	End Time	Duration
1	A dog was walking around. When he came across a butcher, he saw a bone.	00:00:00.630	00:00:10.230	00:00:09.600
2	Hmm, that looked good.	00:00:10.230	00:00:11.990	00:00:01.760
3	He looked around carefully, nobody there.	00:00:12.010	00:00:15.640	00:00:03.630
4	He ran into the shop, took the bone and took off as fast as he could. He ran far awa...	00:00:15.660	00:00:21.910	00:00:06.250
5	There he stopped to catch some breath, the bone in his mouth.	00:00:21.920	00:00:24.850	00:00:02.930
6	Water was streaming under the bridge.	00:00:25.000	00:00:27.830	00:00:02.830
7	The dog, still panting, looked into the water	00:00:27.850	00:00:31.180	00:00:03.330
8	Suddenly he was alarmed by what he saw. There in the water was another dog.	00:00:31.370	00:00:33.810	00:00:02.440
9	The dog barked.	00:00:33.820	00:00:35.280	00:00:01.460
10	The dog saw that the other dog had a bone as well.	00:00:35.280	00:00:38.290	00:00:03.010
11	"That other bone is much bigger, mine is small.	00:00:38.300	00:00:41.290	00:00:02.990
12	If only I had that bone!"	00:00:41.290	00:00:43.470	00:00:02.180

video player

controls

crosshair

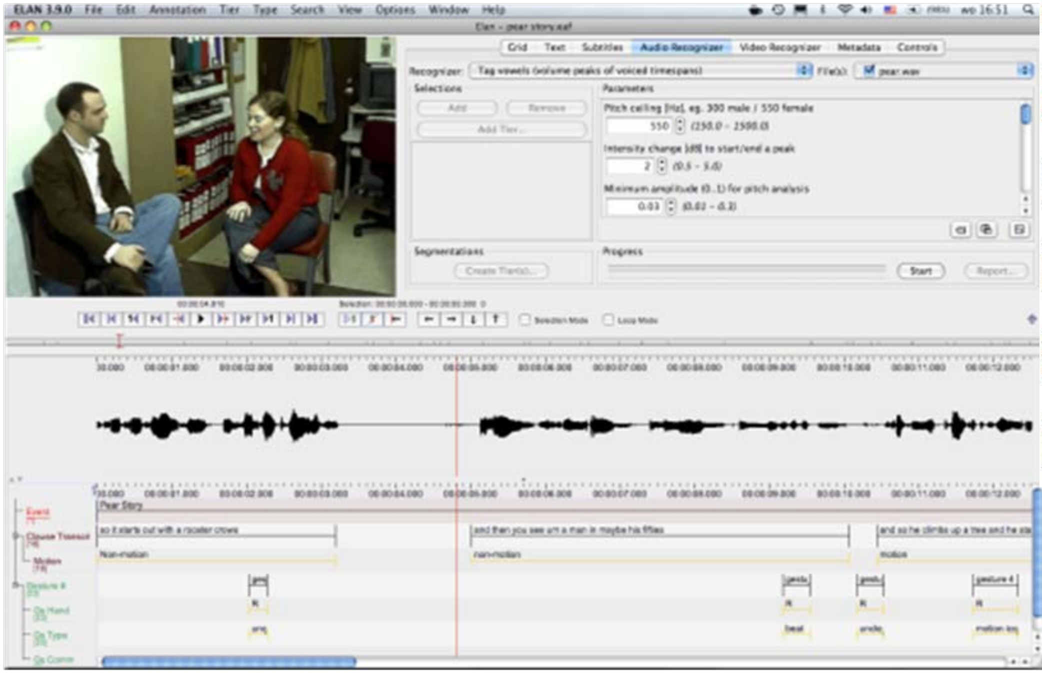
waveform viewer

annotations

timeline viewer

tiers

Computational Methods



Recognizer: Tag vowels (volume peaks of voiced timespans)

File(s): ☑ pear.wav

Parameters

Pitch ceiling [Hz], eg. 300 male / 550 female: 550 (250.0 - 2500.0)

Intensity change [dB] to start/end a peak: 2 (0.5 - 5.0)

Minimum amplitude (0..1) for pitch analysis: 0.03 (0.02 - 0.2)

Pear Story

so it starts out with a rooster crows | Non-motion

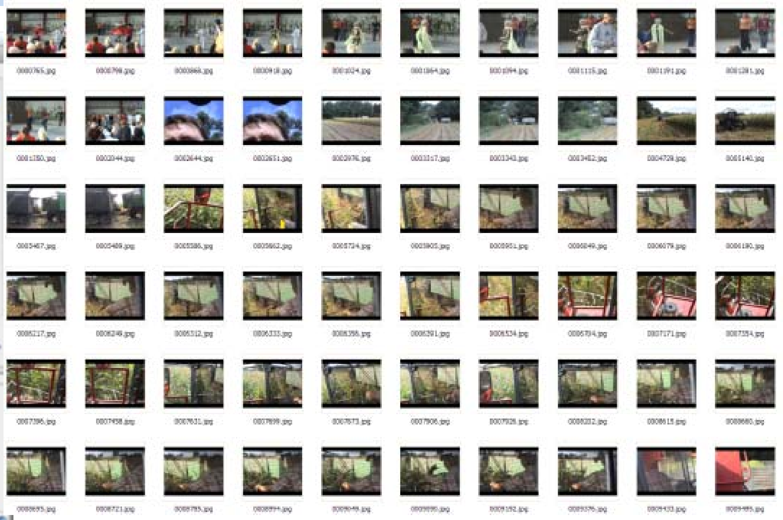
and then you see um a man in maybe his fifties | non-motion

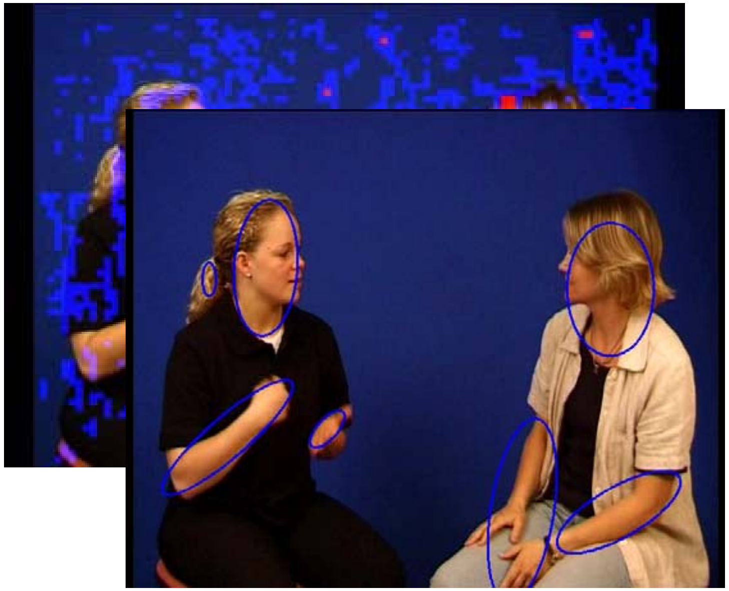
and so he climbs up a tree and he sta | motion

gesture 4 | R | motion iss

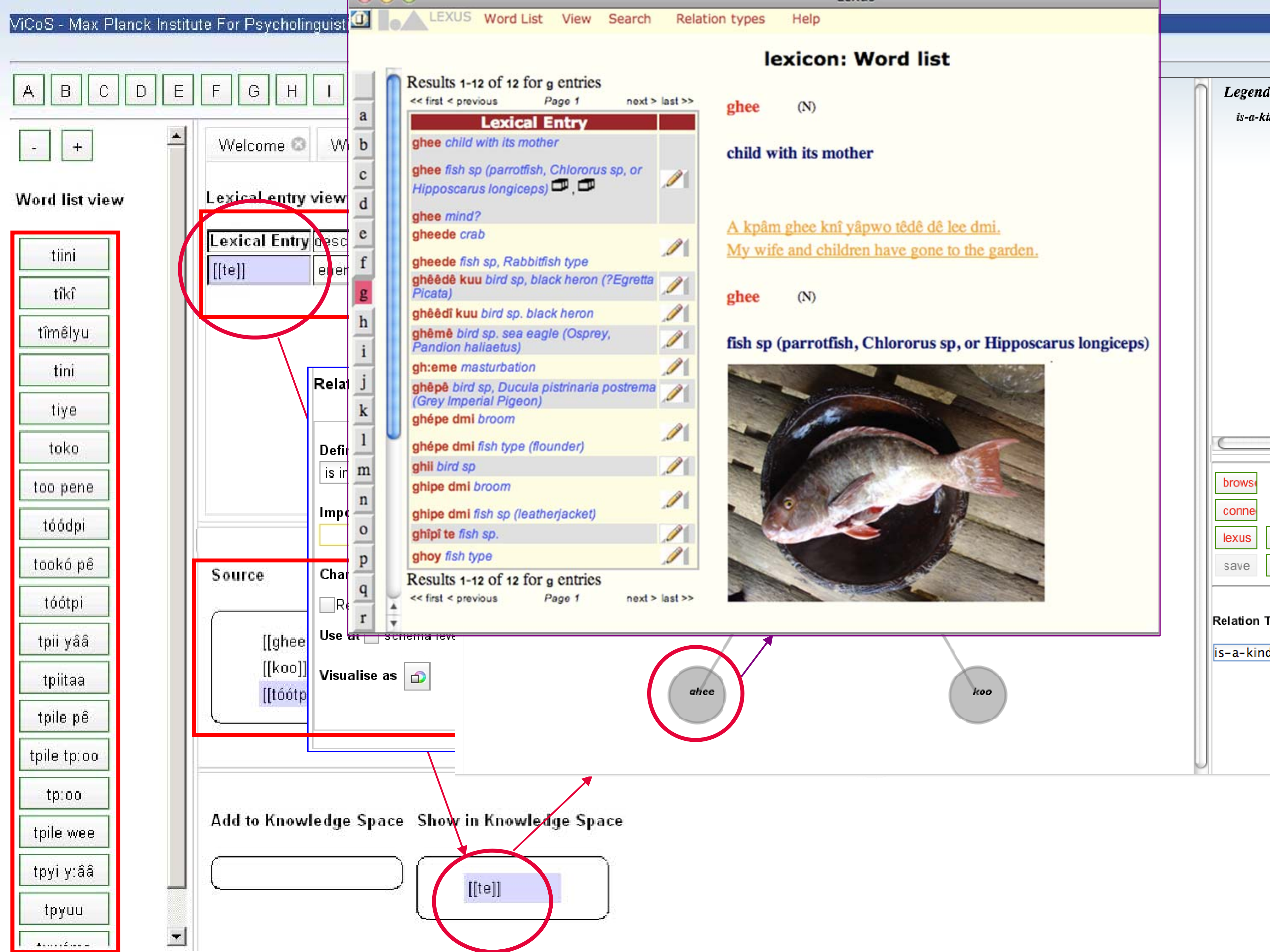
beat

circle





- library of AV detectors to do automatic cumulative annotation
- well help in efficiency increase and in theorization



ViCoS - Max Planck Institute For Psycholinguist

A B C D E F G H I

- +

Word list view

tiini
tîkî
tîmêlyu
tini
tiye
toko
too pene
tóódpi
tookó pê
tóótpi
tpii yââ
tpiitaa
tpile pê
tpile tp:oo
tp:oo
tpile wee
tpyi y:ââ
tpyuu

Welcome W

Lexical entry view

Lexical Entry	desc
[[te]]	ener

Relat

Defi
is ir

Impo

Char

□ Re

Use at □ schema leve

Visualise as

Source

[[ghee
[[koo]]
[[tóótp

Add to Knowledge Space

Show in Knowledge Space

[[te]]

LEXUS Word List View Search Relation types Help

lexicon: Word list

Results 1-12 of 12 for g entries

<< first < previous Page 1 next > last >>

Lexical Entry	
ghee child with its mother ghee fish sp (parrotfish, Chlororus sp, or Hipposcarus longiceps) , ghee mind?	✎
gheede crab gheede fish sp, Rabbitfish type	✎
ghêêdê kuu bird sp, black heron (?Egretta Picata)	✎
ghêêdî kuu bird sp. black heron	✎
ghêmê bird sp. sea eagle (Osprey, Pandion haliaetus)	✎
gh:eme masturbation	✎
ghêpê bird sp, Ducula pistrinaria postrema (Grey Imperial Pigeon)	✎
ghépe dmi broom ghépe dmi fish type (flounder)	✎
ghii bird sp	✎
ghipe dmi broom ghipe dmi fish sp (leatherjacket)	✎
ghîpî te fish sp.	✎
ghoy fish type	✎

Results 1-12 of 12 for g entries

<< first < previous Page 1 next > last >>

a b c d e f g h i j k l m n o p q r

ghee (N)

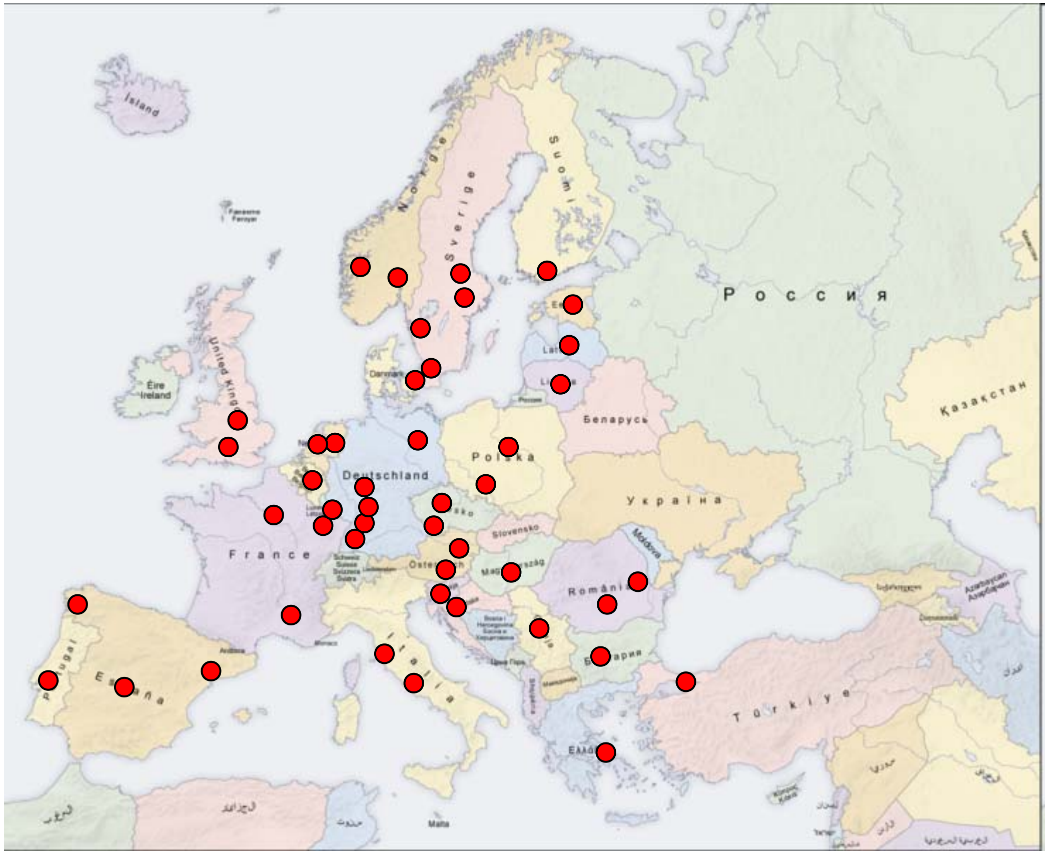
child with its mother

A kpâm ghee knî yâpwo têdê dê lee dmi.
My wife and children have gone to the garden.

ghee (N)

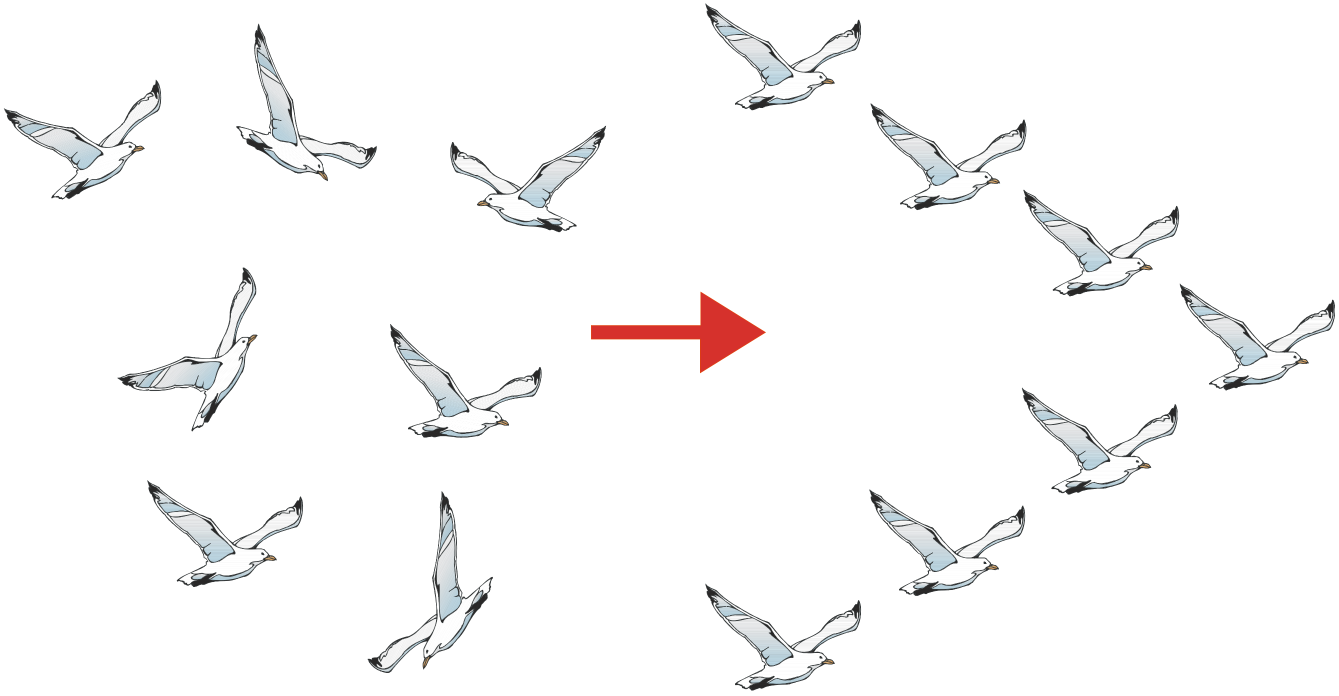
fish sp (parrotfish, Chlororus sp, or Hipposcaros longiceps)

the CLARIN initiative



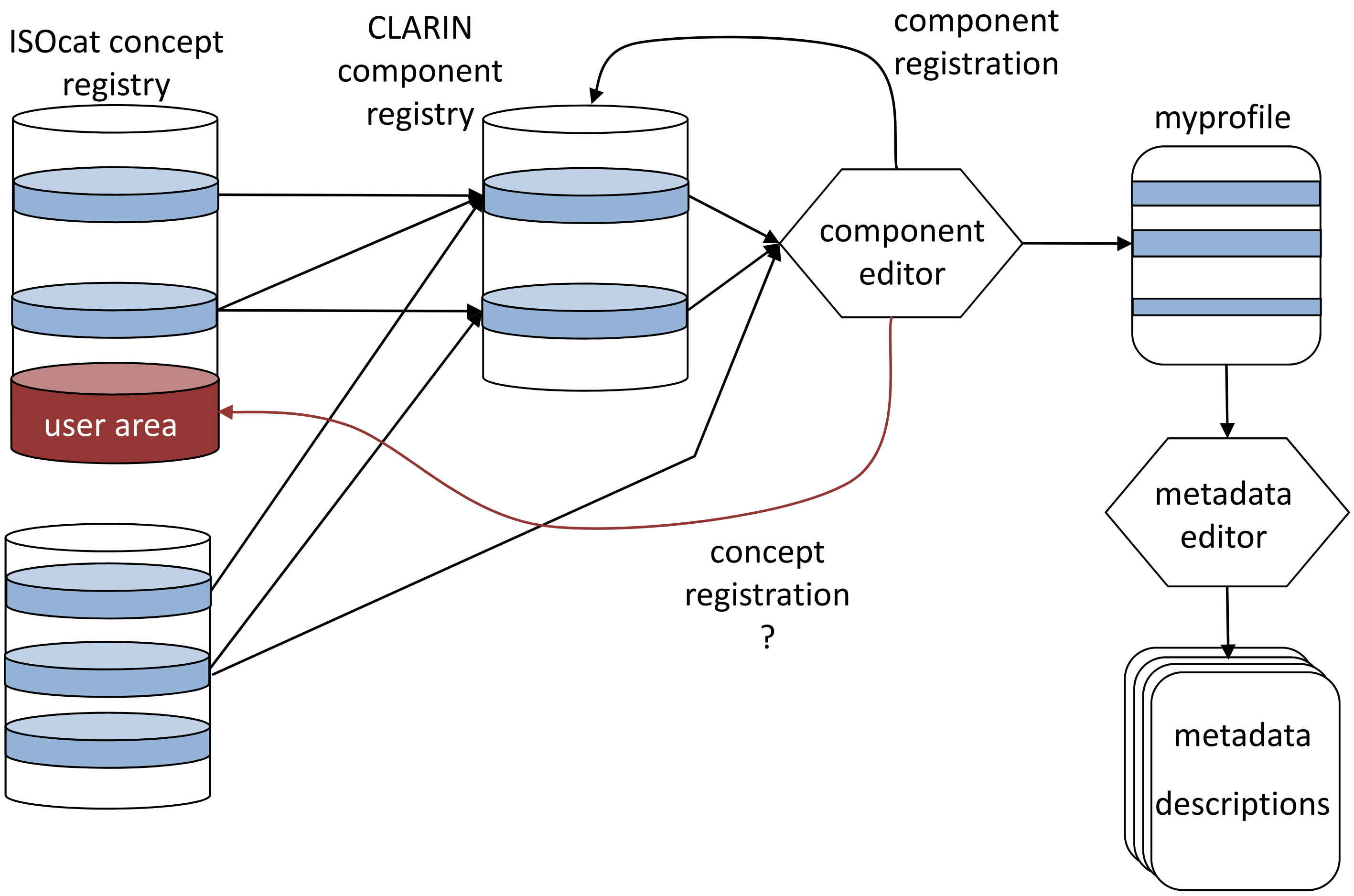
- European research infrastructure initiative funded by EC and member states
- meant to be persistent for many years
- meant to overcome fragmentation in our domain, lack of virtual integration and interoperability
- currently >180 of the most strongest institutes in the field from almost all EU countries
- personal: leader of the technical infrastructure work

Problem to solve in CLARIN

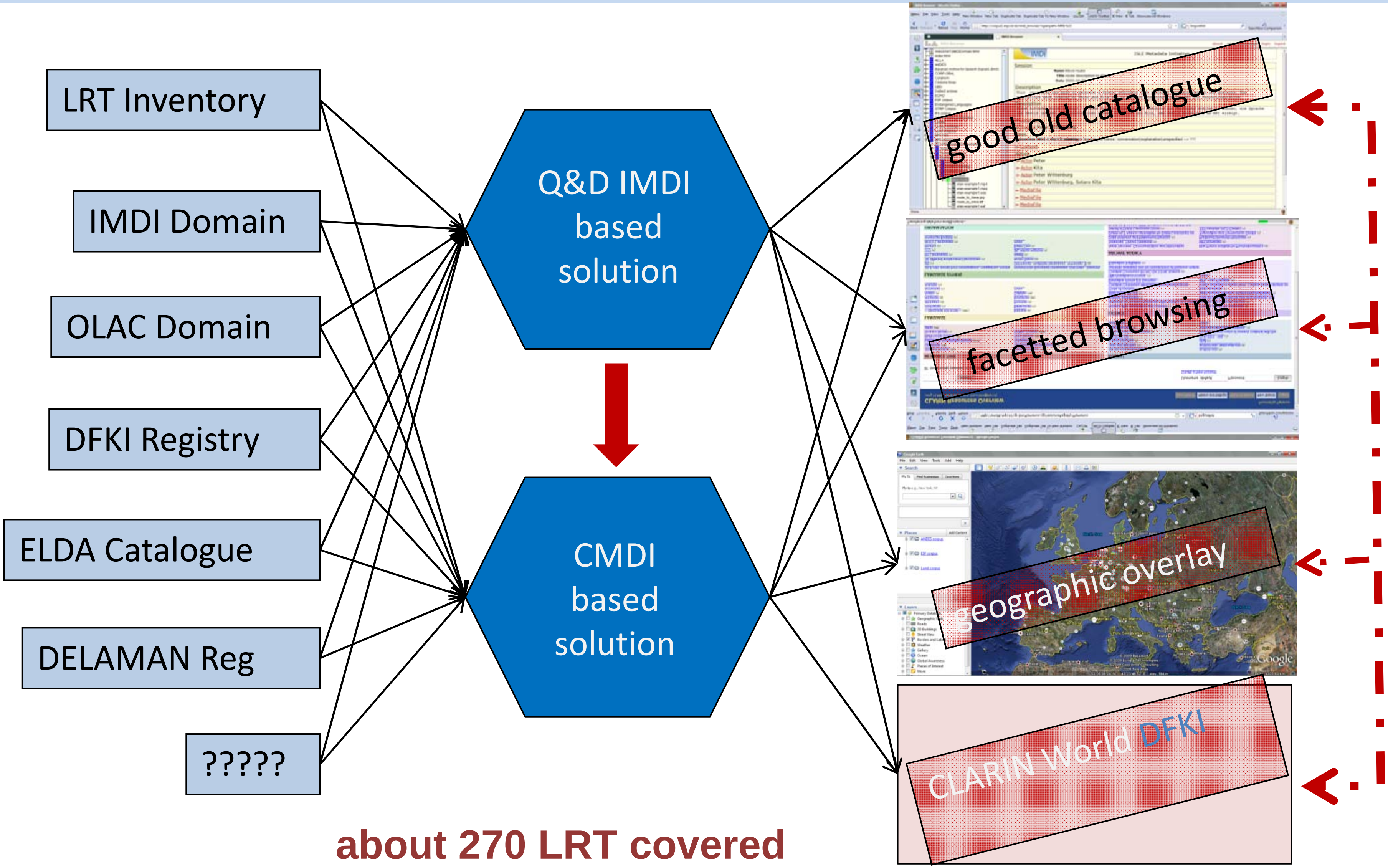


- how to synchronize all minds in our field?
- how much synchronization is good for our field?
- how to synchronize data and tool creation?

Joint Metadata Domain



Virtual Language World



LRT Inventory

IMDI Domain

OLAC Domain

DFKI Registry

ELDA Catalogue

DELAMAN Reg

?????

Q&D IMDI based solution

CMDI based solution

good old catalogue

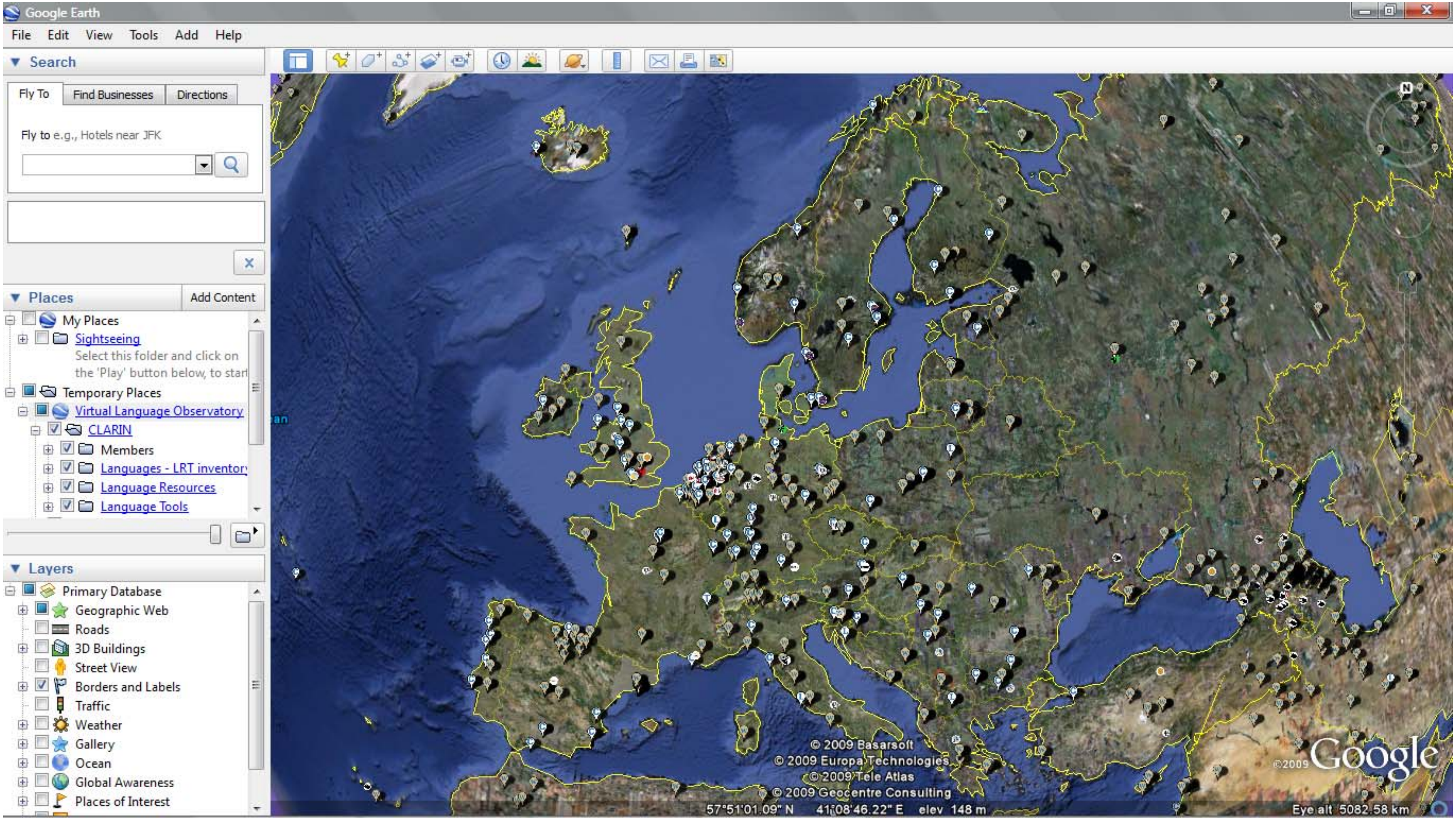
facetted browsing

geographic overlay

CLARIN World DFKI

about 270 LRT covered
www.clarin.eu/vlo

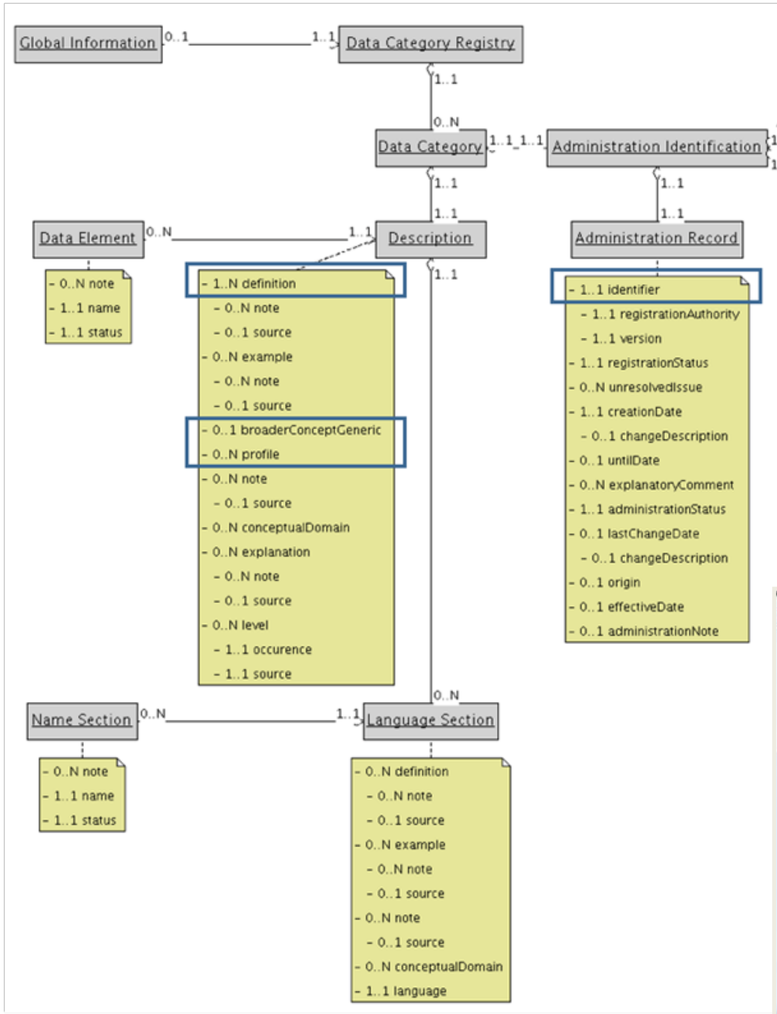
VLO - GoogleEarth Overlay


Google Earth
File Edit View Tools Add Help
Search
Fly To
Find Businesses
Directions
Fly to e.g., Hotels near JFK
Places
Add Content
Layers
an
© 2009 Basarsoft
© 2009 Europa Technologies
© 2009 Tele Atlas
© 2009 Geocentre Consulting
©2009 Google
57°51'01.09" N 41°08'46.22" E elev 148 m
Eye alt 5082.58 km

- □ My Places
  - □ Sightseeing
    Select this folder and click on the 'Play' button below, to start
- ■ Temporary Places
  - ■ Virtual Language Observatory
    - ☑ CLARIN
      - ☑ Members
      - ☑ Languages - LRT inventory
      - ☑ Language Resources
      - ☑ Language Tools

- ■ Primary Database
  - ■ Geographic Web
  - □ Roads
  - □ 3D Buildings
  - □ Street View
  - ☑ Borders and Labels
  - □ Traffic
  - □ Weather
  - □ Gallery
  - □ Ocean
  - □ Global Awareness
  - □ Places of Interest

Semantic Interoperability in Future?

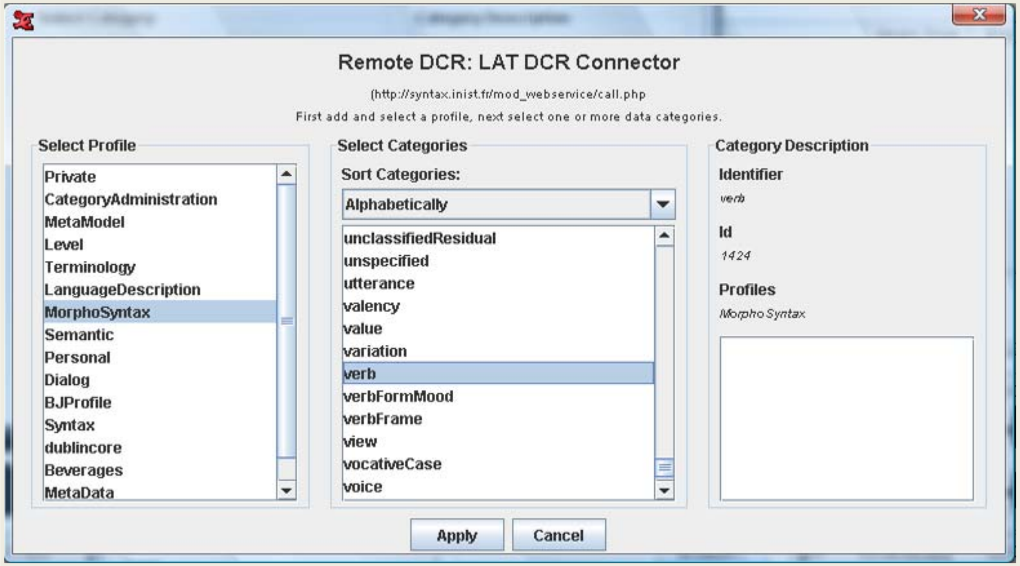


ISO 12620 model

ISOcat implementation

generic model to be used by many disciplines

of course restricted model



Remote DCR: LAT DCR Connector

(http://syntax.inist.fr/mod_webservice/call.php

First add and select a profile, next select one or more data categories.

Select Profile

Private
CategoryAdministration
MetaModel
Level
Terminology
LanguageDescription
MorphoSyntax
Semantic
Personal
Dialog
BJProfile
Syntax
dublincore
Beverages
MetaData

Select Categories

Sort Categories: Alphabetically

unclassifiedResidual
unspecified
utterance
valency
value
variation
verb
verbFormMood
verbFrame
view
vocativeCase
voice

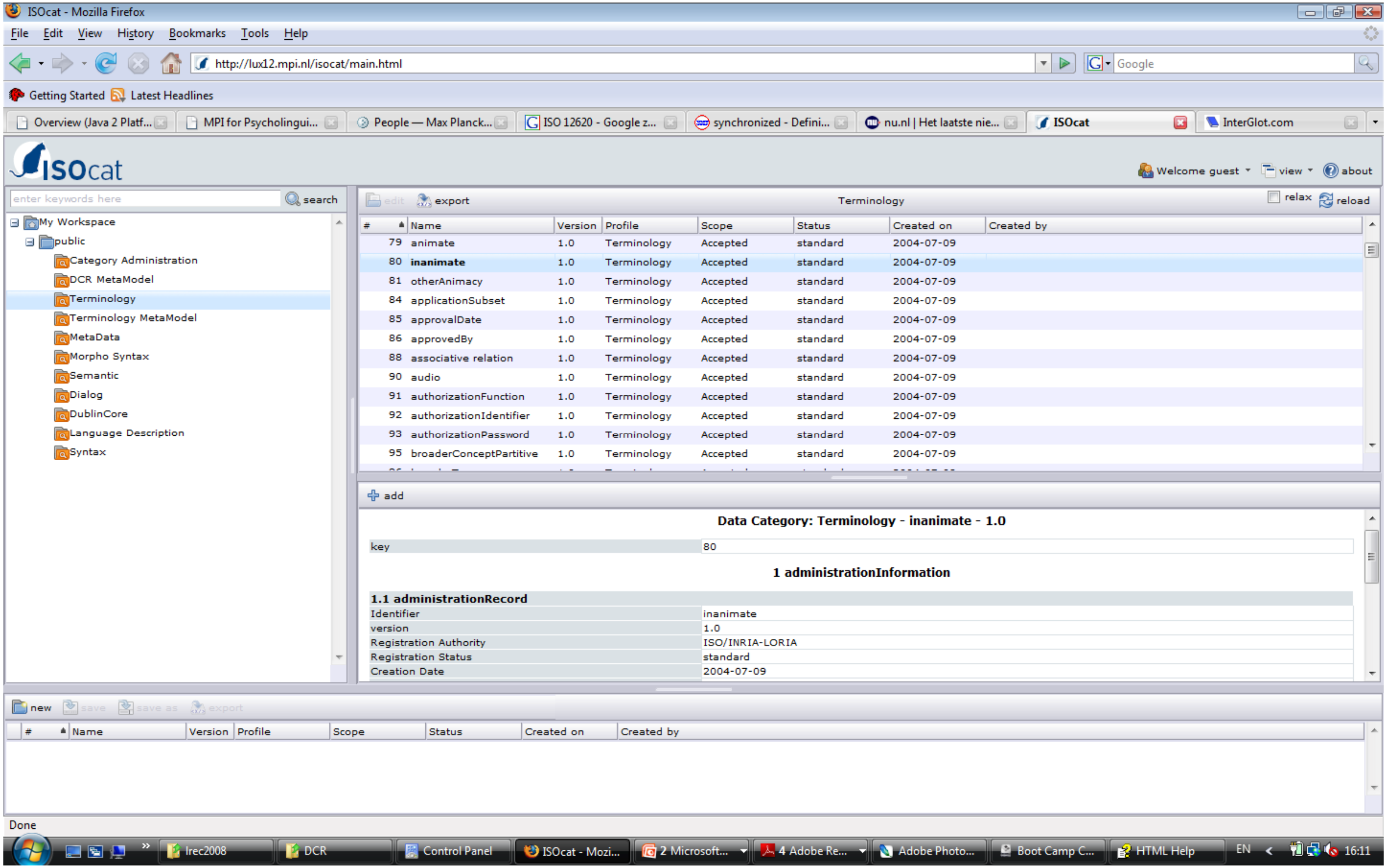
Category Description

Identifier
verb

Id
1424

Profiles
Morpho Syntax

ISOcat is real



- My Workspace
  - public
    - Category Administration
    - DCR MetaModel
    - Terminology
    - Terminology MetaModel
    - MetaData
    - Morpho Syntax
    - Semantic
    - Dialog
    - DublinCore
    - Language Description
    - Syntax

Terminology

#	Name	Version	Profile	Scope	Status	Created on	Created by
79	animate	1.0	Terminology	Accepted	standard	2004-07-09	
80	**inanimate**	1.0	Terminology	Accepted	standard	2004-07-09	
81	otherAnimacy	1.0	Terminology	Accepted	standard	2004-07-09	
84	applicationSubset	1.0	Terminology	Accepted	standard	2004-07-09	
85	approvalDate	1.0	Terminology	Accepted	standard	2004-07-09	
86	approvedBy	1.0	Terminology	Accepted	standard	2004-07-09	
88	associative relation	1.0	Terminology	Accepted	standard	2004-07-09	
90	audio	1.0	Terminology	Accepted	standard	2004-07-09	
91	authorizationFunction	1.0	Terminology	Accepted	standard	2004-07-09	
92	authorizationIdentifier	1.0	Terminology	Accepted	standard	2004-07-09	
93	authorizationPassword	1.0	Terminology	Accepted	standard	2004-07-09	
95	broaderConceptPartitive	1.0	Terminology	Accepted	standard	2004-07-09	

Data Category: Terminology - inanimate - 1.0

key	80

1 administrationInformation

1.1 administrationRecord

Identifier	inanimate
version	1.0
Registration Authority	ISO/INRIA-LORIA
Registration Status	standard
Creation Date	2004-07-09

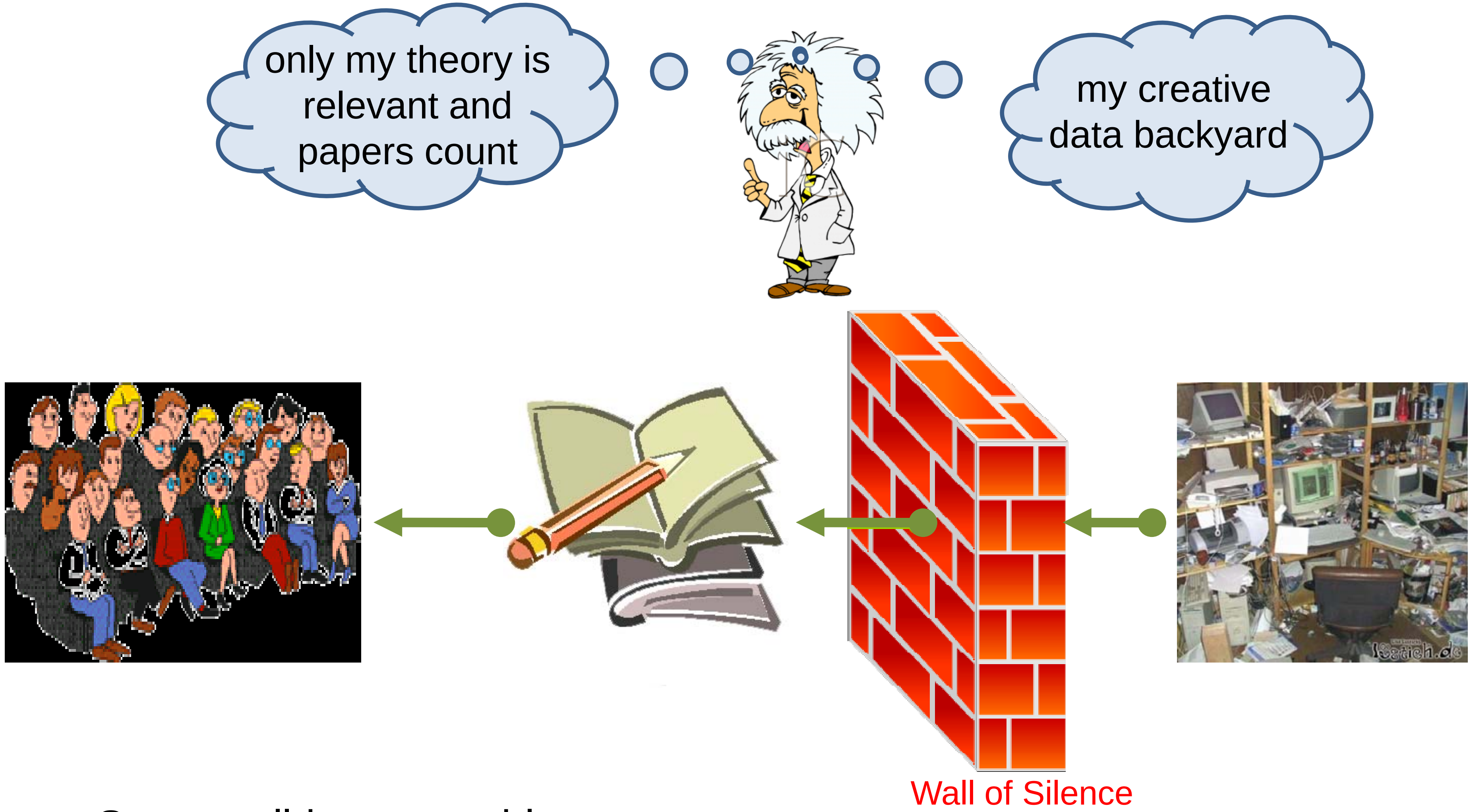
#	Name	Version	Profile	Scope	Status	Created on	Created by

The Data Management Problem

back to the topic of my talk

some major dimensions of concern

Private Data Backyard in SSH



only my theory is relevant and papers count

my creative data backyard

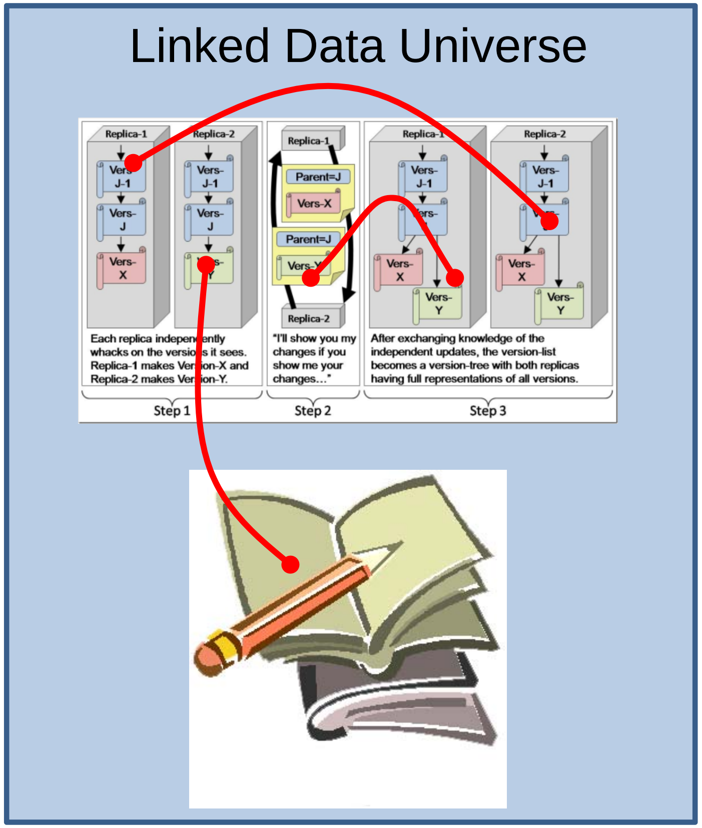
Wall of Silence

Some well-known problems:

no-persistency, hardly any sharing, no correctness proof, etc.

Change in Culture necessary but ...

should I really look into this data mess?



Each replica independently whacks on the versions it sees. Replica-1 makes Version-X and Replica-2 makes Version-Y.

"I'll show you my changes if you show me your changes..."

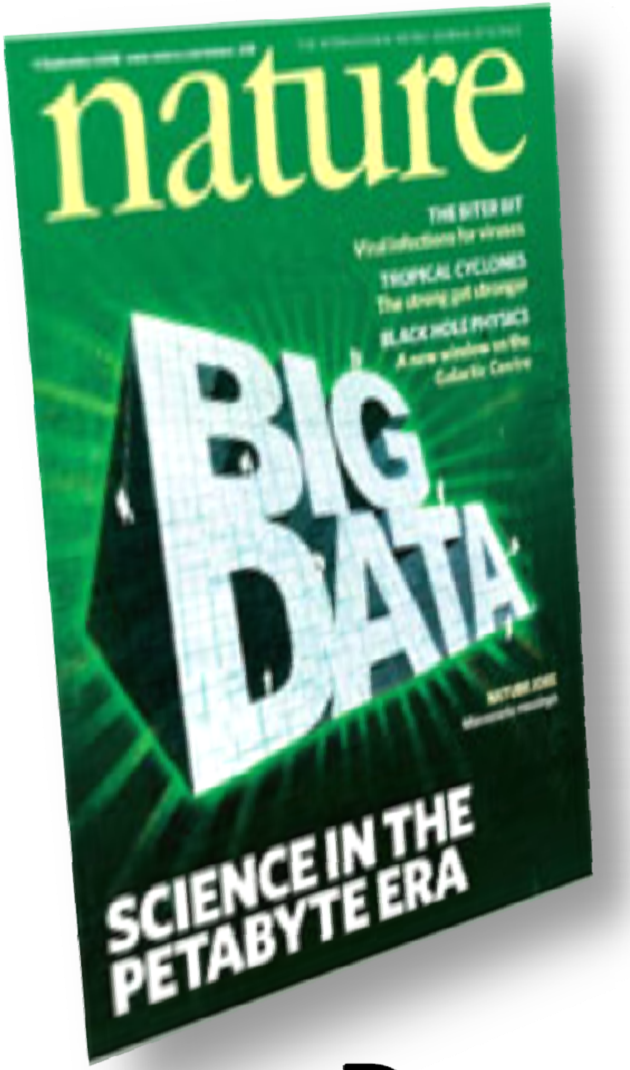
After exchanging knowledge of the independent updates, the version-list becomes a version-tree with both replicas having full representations of all versions.

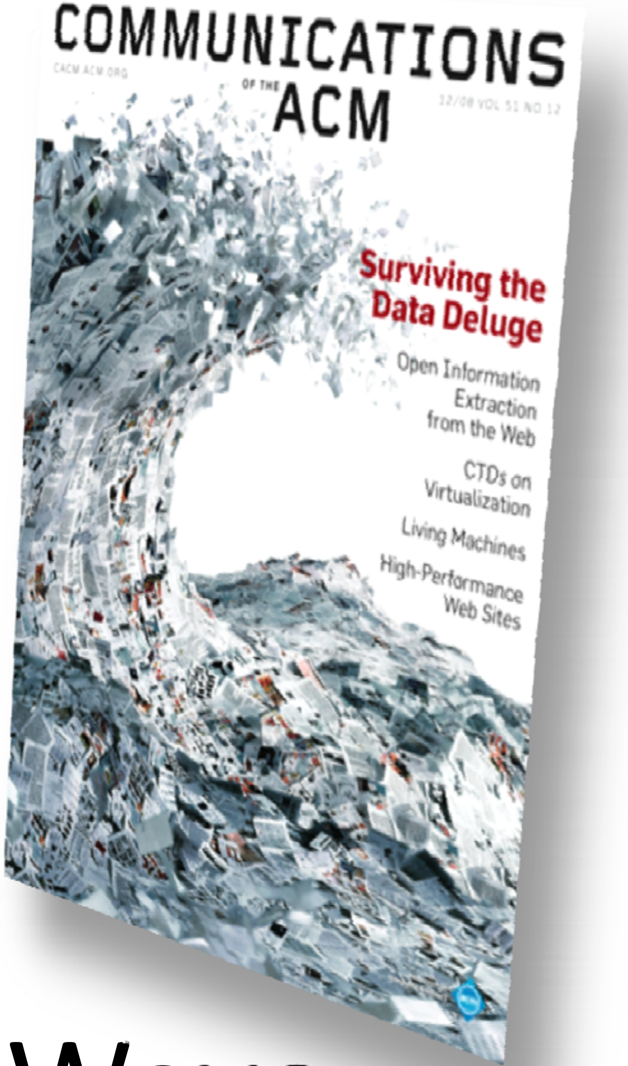
why should I change?

Change in culture required - will not be that easy:

- more work (management, curation), costs?, career?, quality?, etc.
- benefits for small and grand research challenges?

the problem seen by everyone

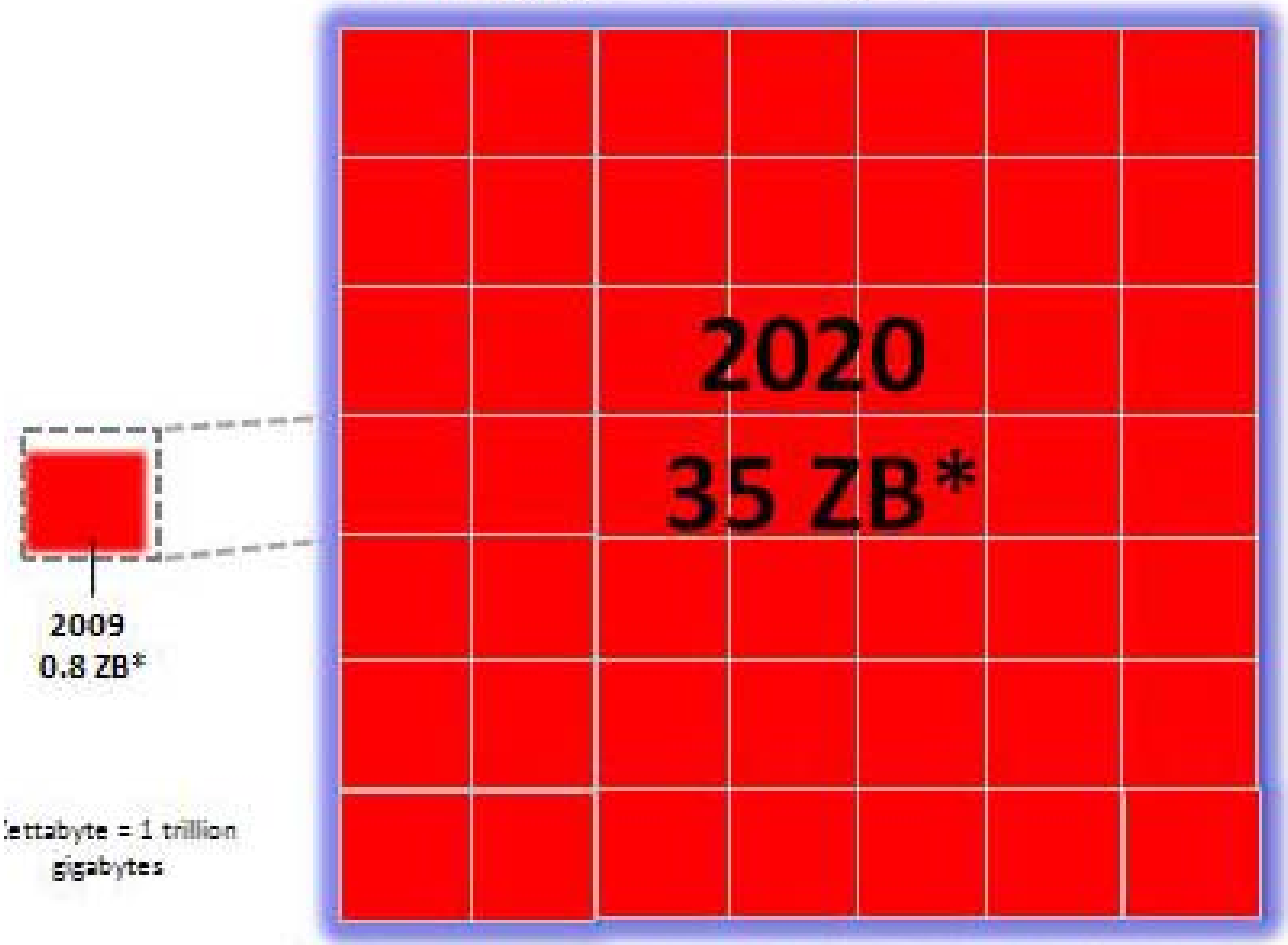

nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
SCIENCE IN THE PETABYTE ERA


COMMUNICATIONS OF THE ACM
CACM.ACM.ORG
12/08 VOL.51 NO.12
Surviving the Data Deluge
Open Information Extraction from the Web
CTDs on Virtualization
Living Machines
High-Performance Web Sites

Data Tidal Wave

Global Data Development

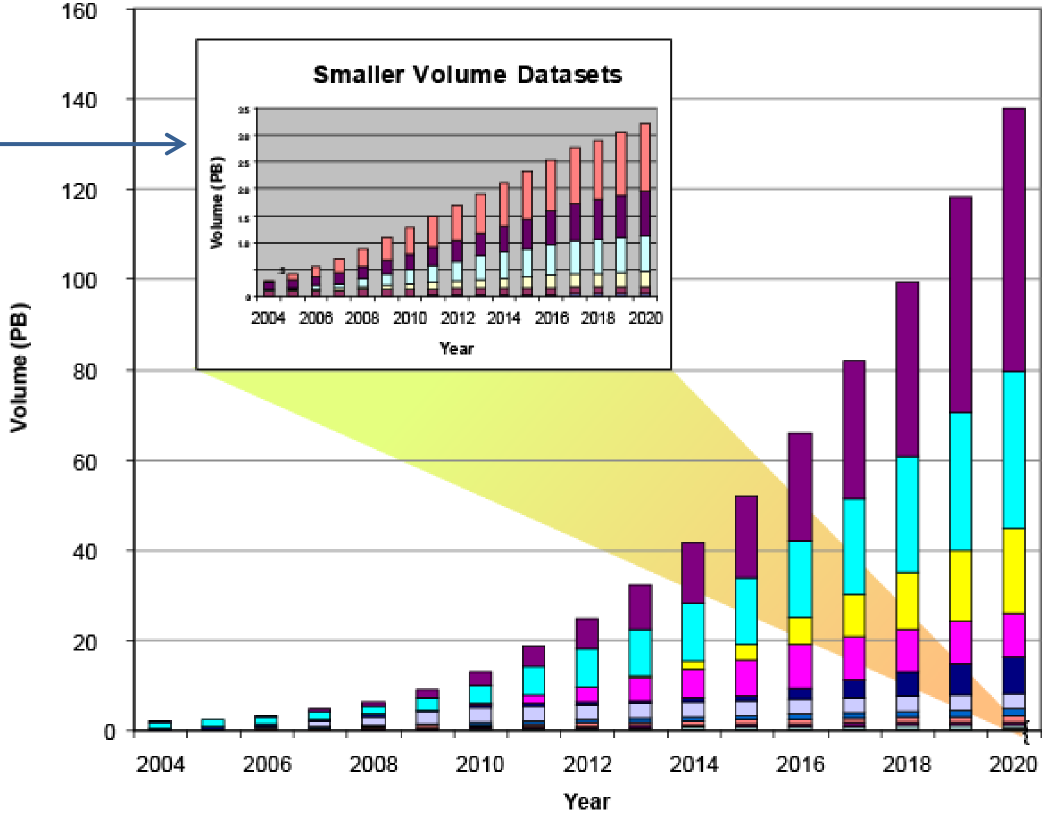
Growing by a Factor of 44



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

US Environmental Data Archive

Comprehensive Large Array-data Stewardship System (CLASS) Storage



Smaller Volume Datasets

Space Based Data

- ■ Polar Orbiting Earth Satellites (POES)
- ■ Defense Meteorological Satellites Program (DMSP)

Earth Based Data

- ■ Atmosphere(Weather & Climate)
- □ Ocean (Weather & Climate)
- □ Continually Operating Reference Stations (CORS)
- ■ Misc (Mesonets)

Large Volume Datasets

Space Based Data

- ■ NOAA Polar-orbiting Operational Environmental Satellite System (NPOESS)
- ■ NPOESS Preparatory Project (NPP)
- ■ Geostationary Operational Environmental Satellites (GOES)
- □ NASA Earth Observing System(Moderate Resolution Spectroradiometer) (EOS MODIS)
- ■ Meteorological Operational Satellite Program (MetOp)

Earth Based Data

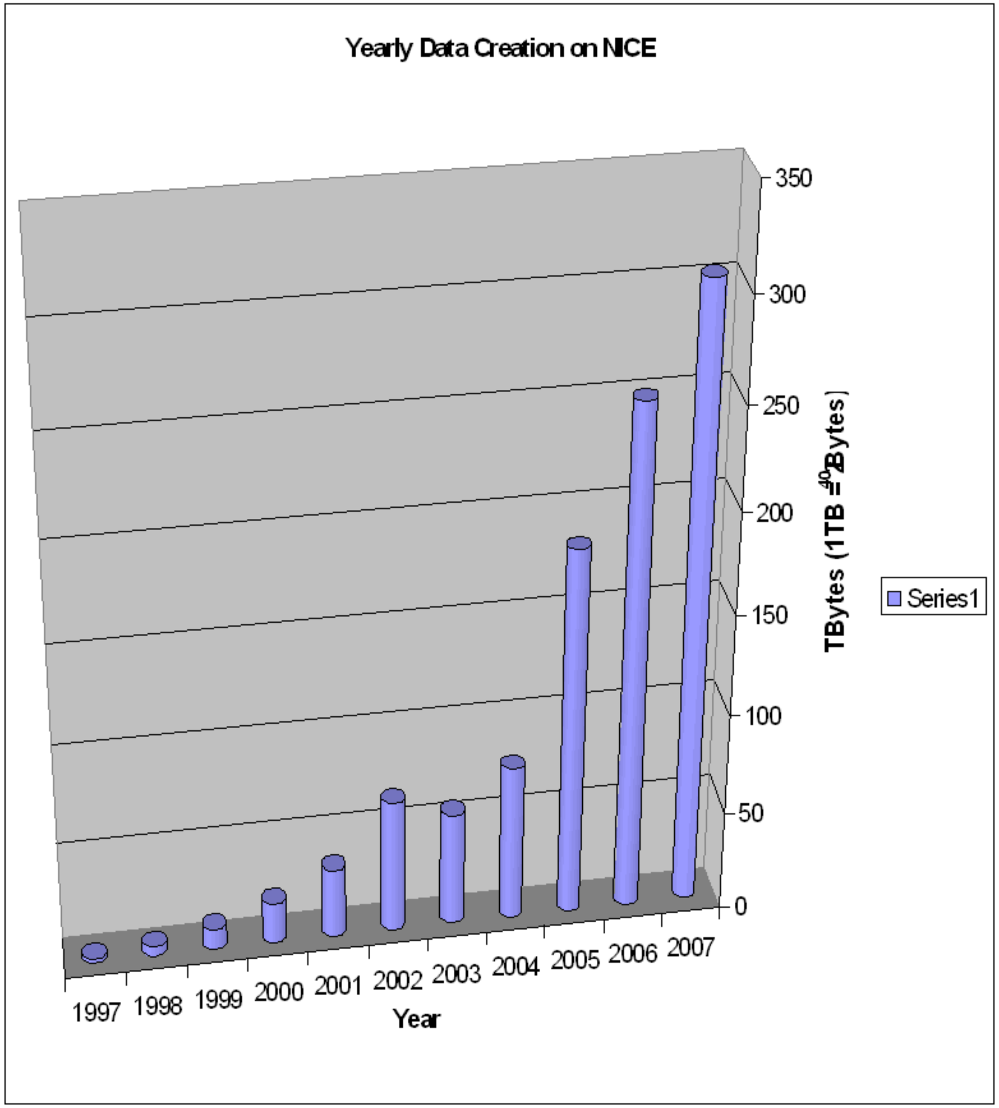
- ■ Weather Radar

Model Data

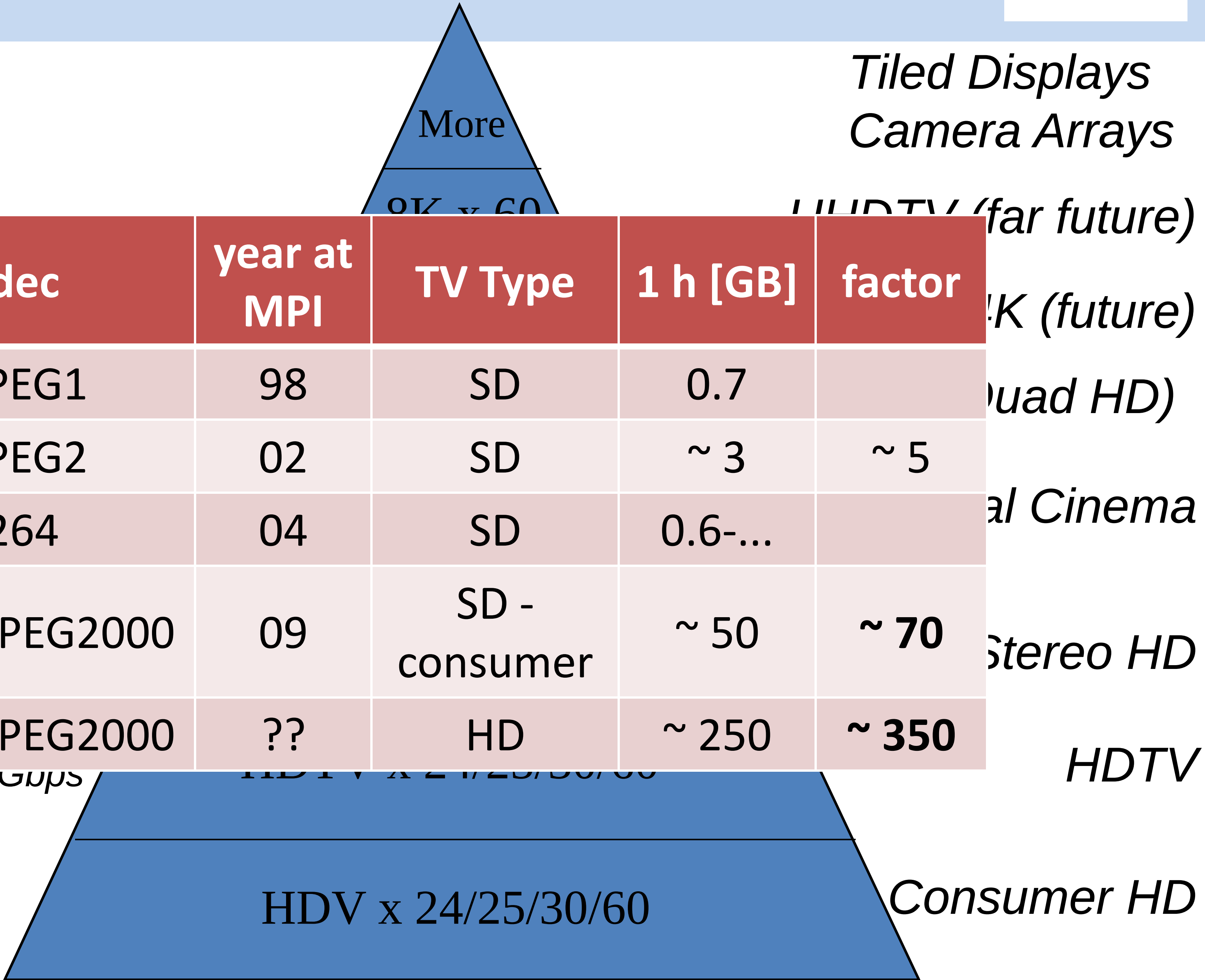
- ■ Atmosphere & Ocean (reanalysis)

European Synchroton Radiation Facility

- 10 years → data volume x 300. In 2007: 300TB ~1*10^8 files
- However, doubling of the data centre infrastructure (m^2, kW, cooling)
- **green computing is an issue**



Video - our SSH domain



Tiled Displays
Camera Arrays

More

8K x 60

1 - 24 Gbp

UHDTV (far future)

500 Mbps

K (future)

250 Mbs -

uad HD)

250 Mbps

al Cinema

200 Mbps

Stereo HD

20 Mbps - 1.5 Gbps

HDTV x 24/25/30/60

HDTV

5 - 25 Mbps

HDV x 24/25/30/60

Consumer HD

codec	year at MPI	TV Type	1 h [GB]	factor
MPEG1	98	SD	0.7	
MPEG2	02	SD	~ 3	~ 5
H.264	04	SD	0.6-...	
mJPEG2000	09	SD - consumer	~ 50	**~ 70**
mJPEG2000	??	HD	~ 250	**~ 350**

Dimensions of Data Management

- thus we see in all disciplines (also in SSH)
  - an enormous increase in shear data volumes
  - AND in number and complexity of resources
- **how can we keep data accessible and usable?**
  - Metadata
  - State of Data
  - Enrichment, Curation and Costs
  - Granularity, Identity and Authenticity
  - Context and Aggregations
  - Preservation and Interpretation
  - Replication and Synchronization
  - Interoperability and Standards
  - Down-Load First and Ecology
  - Sharing, IPR and Quality Assessment

AND Some basic IT Principles

- Atomic Objects
- Explicit Syntax and Declared Semantics
- Persistent Identifiers
- Formats
- Standoff

Relevance of Metadata

- functions of metadata
  - users: find useful resources in increasingly large stacks
  - managers: execute proper algorithms (conversion, IPR, etc)
  - depositors: check state, add new versions, etc
  - funding agencies: value for money
  - communities: what to be preserved

  - machines:
    - harvest them by large portals such as VLO
    - allow smart filtering (virt. collections, community sites)
    - NLP and other chains

Metadata - Change of Culture

- moving towards huge "market" places

- need adequate mechanisms
- need proper tools (immediate metadata creation at lifecycle start such as with cameras)
- requires additional work which is seen as overhead
- but 40 % of researchers' time is spent on finding R&T
- but must become obligatory (NSF, NWO, etc)

State of Data

- tradition in SSH is to include samples in publications as proof and claim that you have the data
- eResearch is different: you need to provide your data

- thus:
  - make it explicit by depositing in a trusted repository
  - will register a unique and persistent identifier (PID)
  - PID will be associated with
    - checksum to proof authenticity
    - time stamp
    - pointers to metadata record
    - pointers to copies

- thus data is citable and identifiable

State - Change of Culture

- thus: we need a deposit culture

- however - there must be a trust relation
  - accessibility, availability, protection, etc
  - could give our data to MS, Amazon, Google, YouTube
  - but then it is "their" data - they can ask money for services
  - do we already have alternatives - JEIN

- however: some linguistic data is never finished or free of bugs
  - lexica, transliterations, annotations, etc
- need a culture of providing imperfect versions and updates

Data Enrichment, Curation & Costs

- unlike traditional publications research data is dynamic
  - new versions, new annotations, new contexts, etc
- thus: any object lives in a context which is changing and which to a large extent is user dependent

- metadata & PIDs at object and collection level can act as glue

- curation (towards proper standards etc) means
  - creating new versions from time to time
  - of course old versions may not be touched
- thus: format/content migration is another source of dynamics

- to maintain interpretability curation is a must
- late curation is very costly
- not coherent collections may not survive due to costs

Curation - Change of Culture

- if long-term accessibility and interpretability is wanted we need to change cultures

- we need
  - support for collection metadata (versions, contexts etc)
  - better support for standards by tools
  - immediate curation at deposit level

- but research is innovative per definition - thus will lack standards
- well - we will always have some chaos somewhere

Granularity, Identity & Authenticity

- identity of an object by an explicitly registered PID record with
  - checksum
  - time stamp
  - pointer to metadata
  - pointer to copies

- but what is an object in linguistics?
- which granularity is appropriate?
  - is it a whole database with all your dynamic data in it?
  - is a container appropriate requiring own application logic?
  - is it a lexical entry which is part of a large lexicon?
- these issues are not at all clear
- at MPI (and others) linguistically meaningful units such as a lexicon, a video, an annotation tier, etc

Identity - Change of Culture

- need a culture of awareness

- a large container is not appropriate since
  - authenticity check etc will depend on own application logic
  - dynamics is too high thus there is no simple versioning
- singular lexical entries are probably not appropriate
  - it makes sense to group them - based on an abstract model
- but what about semantic relations (semantic web)

- every repository needs to define an appropriate solution and make its policies explicit

Context and Aggregations

- contextual information can be of different types
  - metadata including additional information (conditions, etc)
  - source files to understand annotations
  - grouping of thematically related resources
  - hierarchies of such groupings
  - publications associated with a resource
  - etc

- what is the place to manage this information?
  - metadata of recursive type is a good place
  - don't use header information due to a mixture of different information types - against stand-off principles

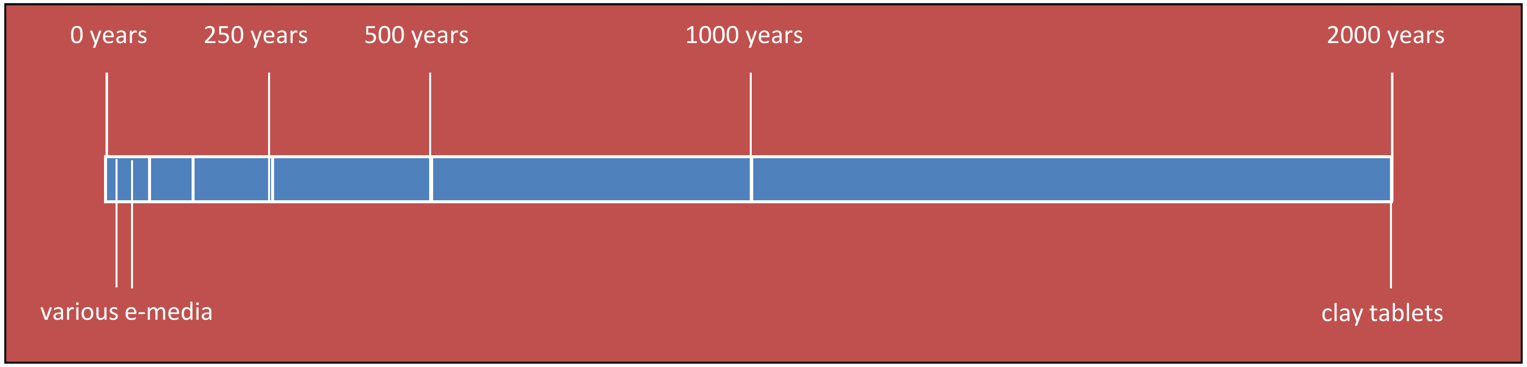
Context - Change of Culture

- obviously need a culture of awareness

- better tools and systematic approach in lifecycle management
- CERIF standard is a candidate for storing all types of non-linguistic information
- stable references are crucial

Preservation, Replication & Interpretation

- bit stream preservation vs. interpretation



- 80% of all language and culture recordings are endangered due to deterioration of carrier substrate
  - for logistic reasons much data will be lost for ever
- same for much of our digital data: formats and encoding standards change frequently
- continuous migration and transformation required
- replication is also required due to vulnerability of carriers, but replication needs to be safe

Preservation - Change of Culture

- need much better and more systematic approaches

- change of Golden Rule:
  - analog: never touch the original and set it aside
  - digital: continuously touch the resources
- digital copying can be lossless - but need checks
- digital migration can be without problems - but need checks
  example: concatenation effects when transforming
  compressed video for example
- replication needs to be done safe and at logical level due to context - currently it is done at physical level with unsafe protocols
- storing provenance information is important to understand the conditions for further processing

Interoperability and Standards

- language resource creation is a highly distributed process
- also linguistic tool development is highly distributed
- the result is a highly fragmented landscape

- is there a Golden Way to overcome interoperability hurdles?

- standards can help to overcome fragmentation
- without question: UNICODE, XML, MPEG, mJPEG2k, lin PCM, language codes (639-3), etc
- huge efforts in ISO TC37 (ISOcat, LMF, LAF, MAF etc), TEI and of course W3C

- but - innovation rate is high and need flexibility for innovation
- converters may help for a while

Interoperability - Change of Culture

- need a change of culture

- software development needs to support emerging standards
- but industry wants to sell tools and does not care about standards from linguists
- linguists and community should be aware and be critical
- community needs to show more discipline - not every new format is really necessary
    - but who decides?
- funders should enforce better behavior

Download first & Ecology

- currently the down-load first paradigm is dominant
  - all data is first downloaded on PC
  - all tools are first installed on PC
- this is
  - very inefficient
  - not ecological due to much uncontrolled copying

- needs to be replaced by cyberinfrastructure
- but yet not enough reliability, availability, security, protection
- industry offers are around - but at cost of rights

- need more interest and benefit for sharing
- need other IPR and trust basis

Download-First - Change of Culture

- need a change of culture

- need reliable and available infrastructures for researchers
- need professional workspaces at centers

- need a mentality of sharing data
- need a complete change in IPR: academic use

Basic IT Principles

- all needs to be based on a few generic IT principles

- create and maintain atomic objects
  - don't mix different types of information
  - earlier granularity discussion
- use explicit syntax
  - obvious but still not common practice and not tool supported
- declare semantics
  - define the concepts you are using
  - ISOcat is a start - will it really take up?
- use PIDs
  - register all references explicitly
  - cool URIs may work for some purposes
- use stand-off principles

Summary

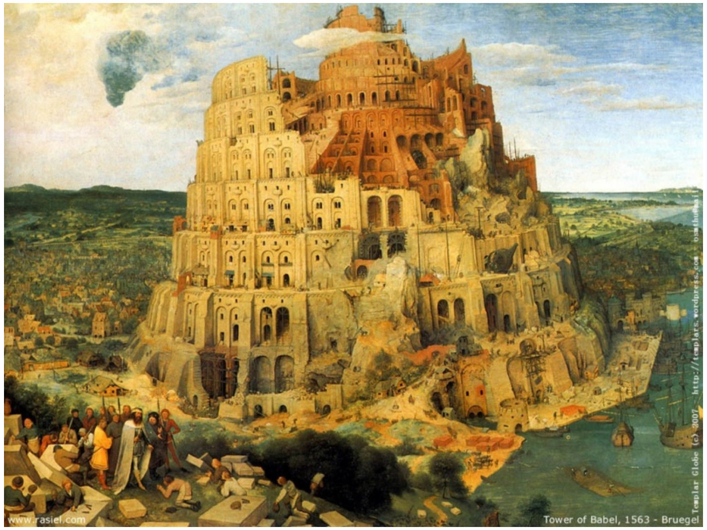
- data driven research is dependent on easy access to all data
- access patterns cannot be predicted due to interdisciplinary research
- technology innovation allows us to create huge amounts and we are using all options
- this creates the need for much better data management strategies

- discussed some dimensions and a few basic principles
- is the situation hopeless?
- no - many ingredients are in place
- but not systematically available and supported
- eResearch is about offering systematic approaches
- CLARIN is one of the initiatives to improve the situation

End

Falls nicht to end in Babylonish scenario nous avons still etwas time üm na te think.

Thanks for your attention!


www.rasiel.com
Tower of Babel, 1563 - Bruegel
Templar Globe (c) 2007 - http://templars.wordpress.com - osmthu@mail