CLARIN
Common Language Resources and Technology Infrastructure

Persistent Identifier Service for CLARIN

Daan Broeder, MPI for Psycholinguistics Nijmegen
Ulrich Schwardmann, GWDG Goettingen

NEERI, Helsinki Oct 1,2

Research Infrastructure for Linguistics and SSH

Research in Linguistics & SSH is increasingly often data driven

- the fuel for this is primary and secondary data – in ever increasing amounts

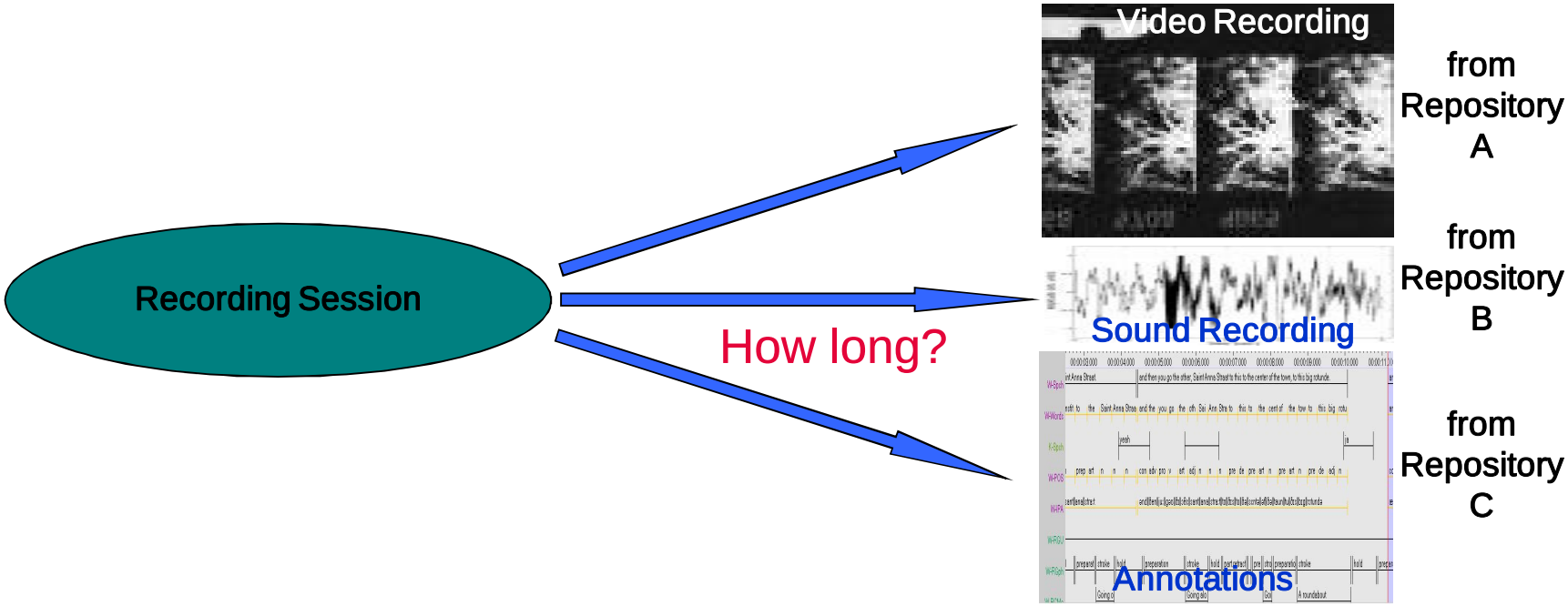
- data is stored in various repositories, so we need virtual collections and reproducibility

- need to refer to relevant data, need to link between data elements to preserve relevant findings, need to link between documents and data elements etc
- thus: one essential pillar for us are persistent references

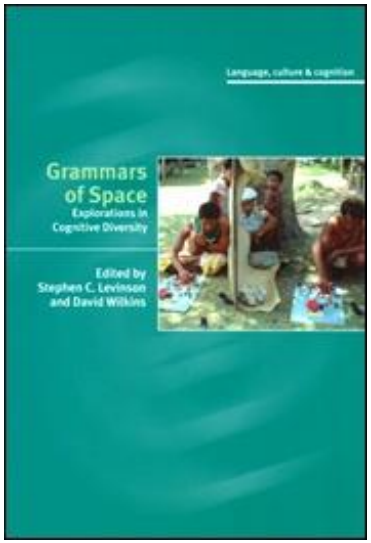
A few examples from our domain

A simple example

- assume that we have a recording of a dying language and some annotations that tell us what someone said about medicine etc
- researchers create relations that need to be preserved

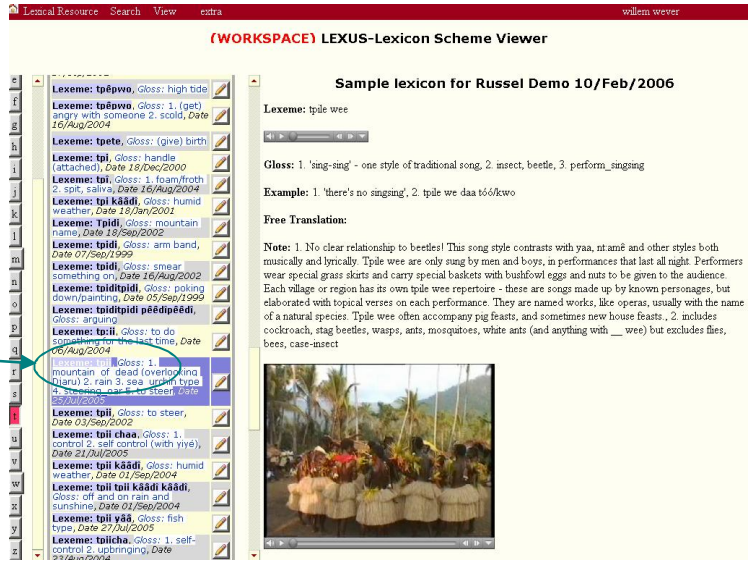


Another simple example



Biological and cultural processes have evolved together, in a symbiotic spiral; they are now indissolubly linked, with human survival unlikely without such culturally produced aids as clothing, cooked food, and tools. The twelve original essays collected in this volume take an evolutionary perspective on human culture, examining the emergence of culture in evolution and the underlying role of brain and cognition. The essay authors, all internationally prominent researchers in their fields, draw on the cognitive sciences -- including linguistics, developmental psychology, and cognition -- to develop conceptual and methodological tools for understanding the interaction of culture and genome. They go beyond the "how" -- the questions of behavioral mechanisms -- to address the "why" -- the evolutionary origin of our psychological functioning. What was the "X-factor," the magic ingredient of culture -- the element that took humans out of the general run of mammals and other highly social organisms?

Several essays identify specific behavioral and functional factors that could account for human culture, including the capacity for "mind reading" that underlies social and cultural learning and the nature of morality and inhibitions, while others emphasize multiple partially independent factors -- planning, technology, learning, and language. The X-factor, these essays suggest, is a set of cognitive adaptations for culture.

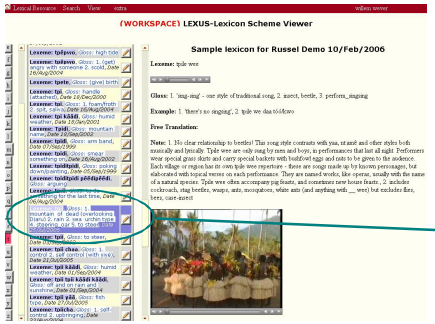


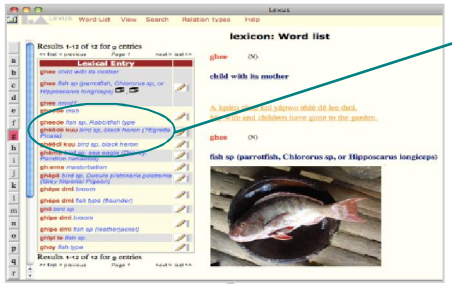
ePublication
Repository 1

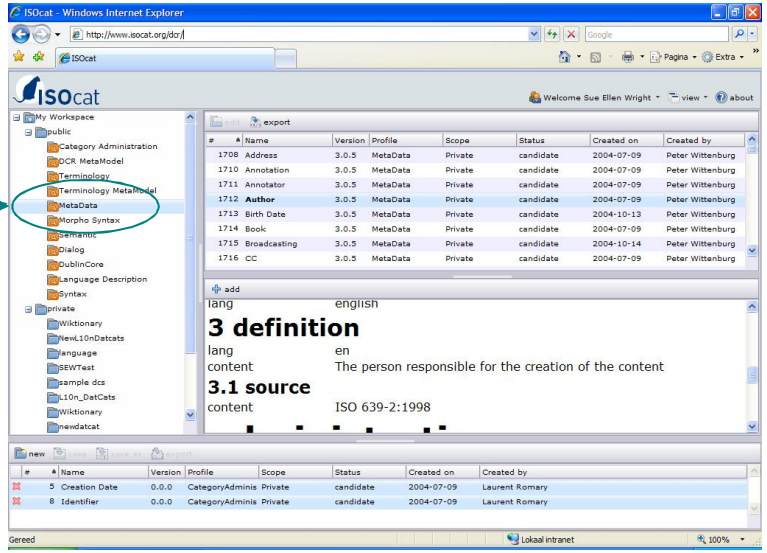
How long?

eResource
Repository 2

Another example







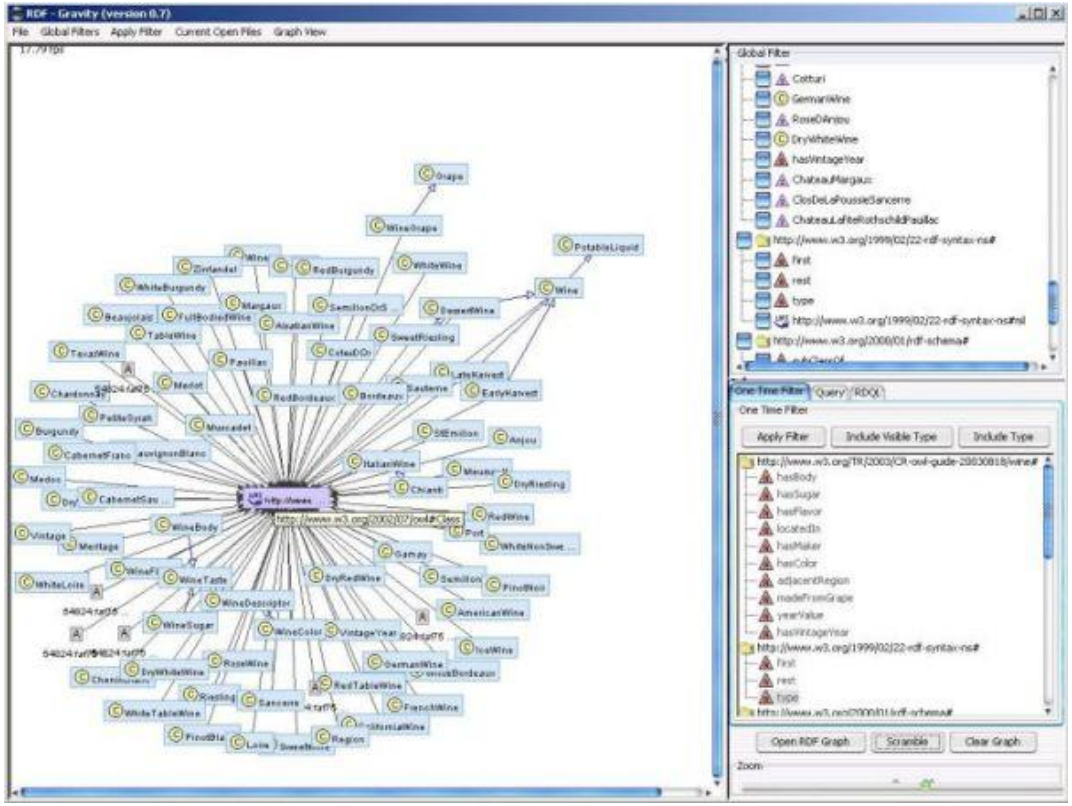
eResource1
Repository 1

eResource2
Repository 2

How long?

Concept
registry

Still another example



eResource

How long?

RDF Assertion

eResource

PIDs are essential but …

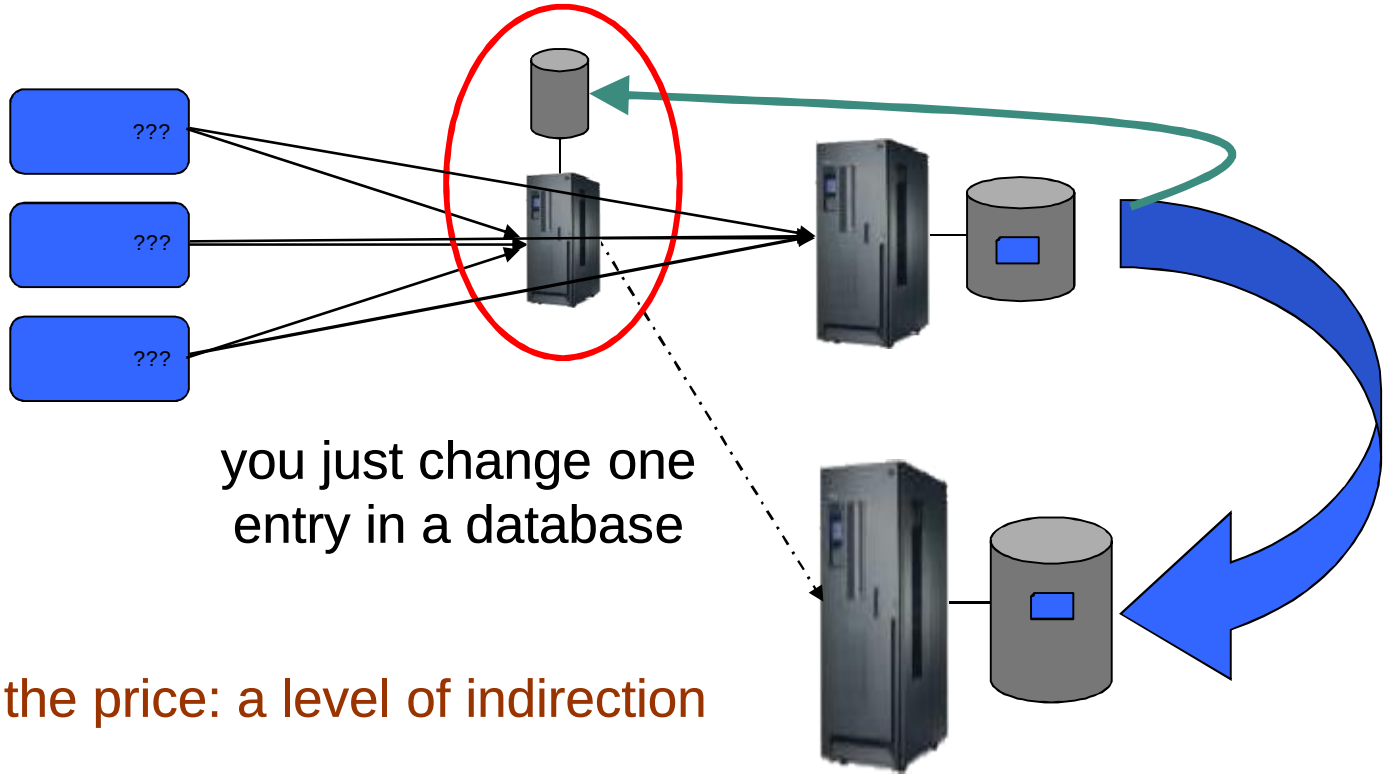
- thus for any research infrastructure the introduction and the use of persistent identifiers is crucial in many ways
- they are so crucial that
  - we need to rely on the fact that they are persistent, thus we must not play anymore
  - research may not be dependent on a commercial service that could develop into a monopoly situation
  - we need to ensure that research is defining the rules of the game

- but there is a price which some (for example W3C TAG) do not want to pay

Is there a problem?

- could use Cool URIs as the W3C TAG suggests to do
- addresses change too often and we cannot influence that
- perhaps some exceptions such as

  http://www.isocat.org/datcat/DC-1708



this is the price: a level of indirection

CLARIN defines some criteria

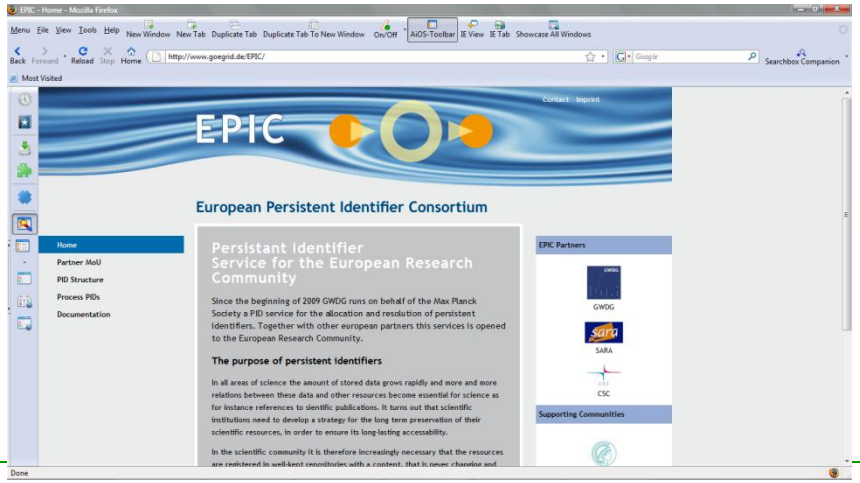
- actionable (URLified) PIDs
- no forced semantics or non-semantics in the PID
- support for fragment addressing
- support of multiple copies
- support for additional (metada) data
- performance, robustness, availability of the service
- secure administration
- sufficiently easy management – suitable APIs
- easy transfer of administration to other organization
- independence from any single service provider
- no single point of failure
- feasible cost model

Ideas for additional services

- want to check authenticity before copying etc (MD5 field)
- want to add citation data info extracted from metadata records
- want to solve the problem of having several centers safely sharing PID information
- want a hook to access permission information
- proper monitoring services

- have a group of centres to support the PID service: the EPIC consortium was born!!!

SARA: the large CC in NL
CSC: the large CC in FI
GWDG: one of the 2 large CC of MPG
look for further strong partners



Choice

We MPI PL, MPG, CLARIN, … made a choice for the Handle System as the basis for a PID service.

Has been a process of years, organizing seminars, making evaluations, talk to competing frameworks, argue with W3C, MPG IT council discussions, having hands-on experience

But other opinions exist, therefore:

- NO ONE is obliged to register PIDs with the CLARIN recommended system
- but CLARIN centres need to be sure that
    - what you do is robust and persistent
    - it can fulfil the essential requirements
    - it can resolve your PIDs☺

- frequently we say: why not register two PIDs ?

Additional issues

- Political Independence:
  - European GHR mirror & proxy + no single point of failure
  - Handle Policy Board
  - Patent ownership
  - Domain ownership

- Wide(r) acceptance of PID scheme by W3C (?)

- Support for object part addressing from ISO TC37/SC4 CITER standard
- Support for multiple administrative sites

How does it work

- finally you need to know
    - what the principle mechanisms are (-> Larry's talk)
    - how you can register PIDs
    - how you can resolve PIDs
    - how you can administer PIDs

See Ulrich's slides and demo

What does the user see?

IMDI-Browser

about user: anonymous login logout

- IMDI-corpora
  - WelcomeToIMDIDomain.html
  - index.html
  - AILLA
  - ANDES
  - Bavarian Archive for Speech Signals (BAS)
  - CGN en IFA Corpora
  - CORP-ORAL
  - Coralrom
  - Corpora Soas
  - DBD
    - Conditions_for_use_DBD.pdf
    - Conditions_for_use_DBD.txt
    - dbd
    - Berber
      - E-rramdani
      - El Aissati
        - tcult (2000)
          - Morocco
          - The Netherlands
            - Utrecht (Lombok)
              - DBD_RIF_14_06_07_
                - nordin_B.cha
              - DBD_RIF_14_09_07_
              - DBD_RIF_14_09_07_
              - DBD_RIF_14_09_07_
              - DBD_RIF_14_09_07_
              - DBD_RIF_14_09_07_
              - DBD_RIF_14_09_07_
              - DBD_RIF_14_12_07_
              - DBD_RIF_14_12_07_
        - DBD_RIF_14_12_01_064
    - Dutch Etnolects
    - Moluccan Malay
    - Moroccan Arabic
    - Papiamentu
    - Russian
    - Sranan:Sarnami
    - Test corpus
    - Turkish
  - DoBeS archive
  - ECHO
  - ESF corpus
  - Endangered Languages
  - GTRP Corpus

Bookmark an archive node or resource

You selected a node from the archive:
name: nordin_B.cha
title: Resource "nordin_B.cha" from "nordin"
internal nodeid: MPI521258#
handle: 1839/00-0000-0000-0007-F42A-6
url: http://corpus1.mpi.nl/qfs1/media-archive/dbd_data/Aissati/tcult_2000_/The_Netherlands/Utrecht_Lombok_/Annotations/nordin_B.cha

To save a reference to the IMDI Browser with the tree opened to this node, bookmark the following link:
nordin_B.cha

Some browsers support direct use of "handle URIs", if that is the case you can bookmark the following link:
Handle link to resource: nordin_B.cha

To save a link to the resource using the central handle resolver, bookmark the following link:
Handle link via central resolver to resource: nordin_B.cha

To save the URL for this resource which is NOT guaranteed to be persistent, bookmark the following link:
URL link to resource: nordin_B.cha

Cite an archive node or resource

You can create citations for this resource by embedding the following text in your document:
Title: Resource "nordin_B.cha" from "nordin"
Author(s): ,
Archive: unknown
Resource format: text/x-chat
Published: Tue Apr 17 14:25:05 CEST 2007
Reference: 1839/00-0000-0000-0007-F42A-6
Link: http://hdl.handle.net/1839/00-0000-0000-0007-F42A-6

In IMDI world 1 Mio PIDs


```
<?xml version="1.0" encoding="UTF-8"?>
<METATRANSCRIPT ArchiveHandle="hdl:1839/00-0000-0000-0005-82B0-2"
    Date="2006-07-18" FormatId="IMDI 3.0"
    Originator="Editor - Profile:SESSION.Profile.xml" Type="SESSION" Version="1"
     xmlns="http://www.mpi.nl/IMDI/Schema/IMDI"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI ./IMDI_3.0.xsd">
 <Session>
       <Name>DBD_RIF_14_12_01_064</Name>
       <Title>Dutch Bilingualism Database, Ethnic Dutch, Session 64</Title>
      …….

<MediaFile>
<ResourceLink ArchiveHandle="hdl:1839/00-0000-0000-0004-DC6B-0">
     http://corpus1.mpi.nl/qfs1/media-archive/dbd_data/boumans/T-
     Cult/Metadata/../Media/dbd_rif_14_12_01_064.wav</ResourceLink>
      …….
```

End

CLARIN will provide further training in the use of the PIDs and the EPIC PID service. Probably at the Prague centre workshop 5,6 Nov

Thanks for your attention