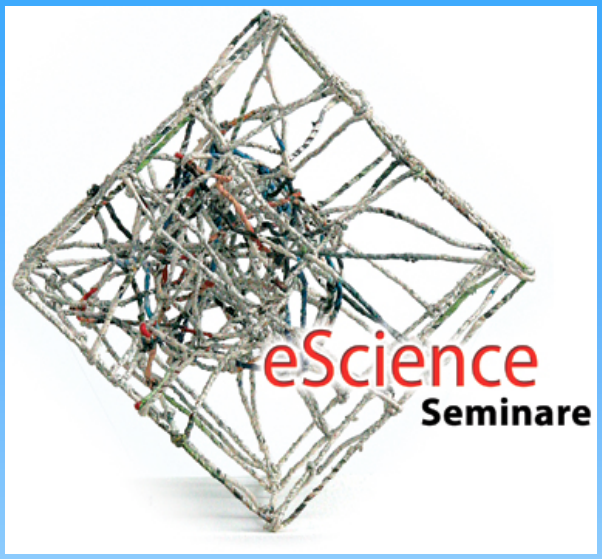


4. MPG eScience Seminar
Unique and Persistent Identifiers

Peter Wittenburg

MPI für Psycholinguistik

CLARIN Research Initiative

Welcome

welcome to the 4th MPG eScience Seminar

fully in line with the first 3 seminars:

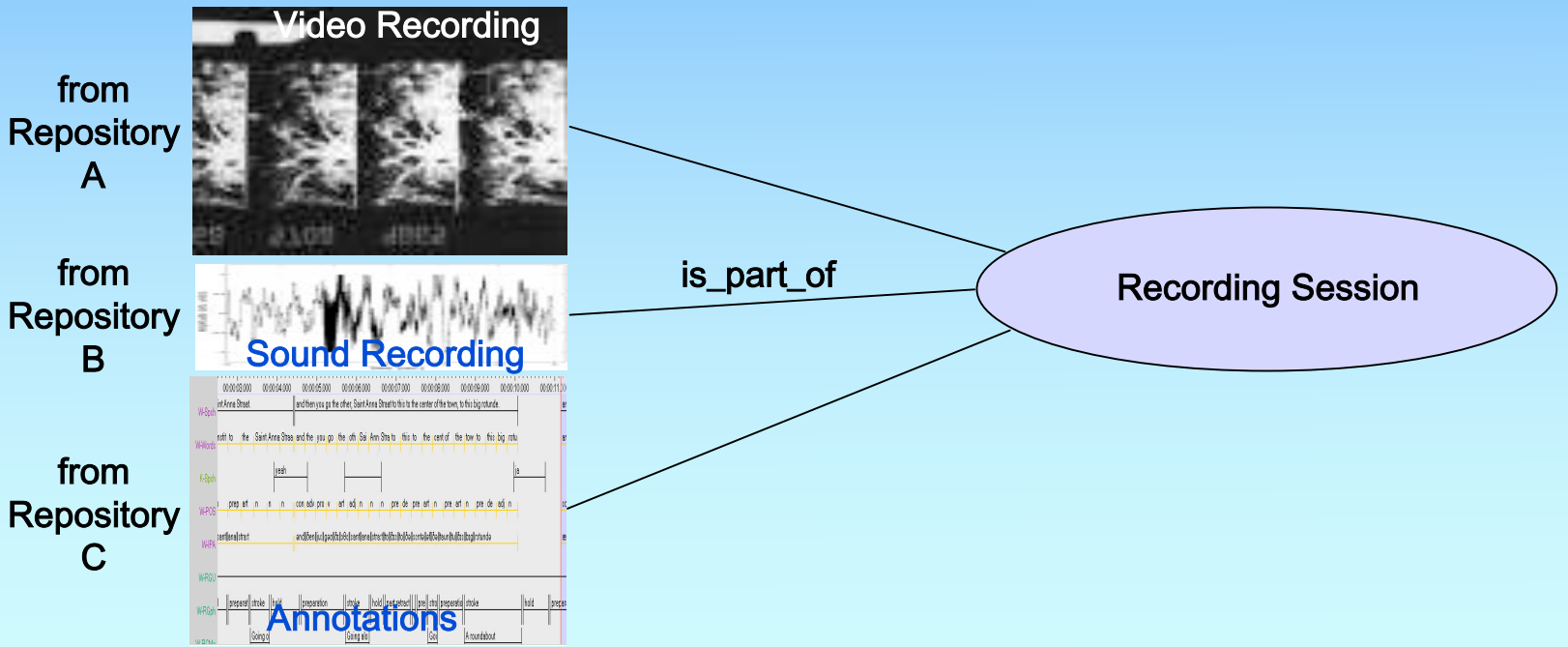
- no abstract discussion of what eScience could mean
  - but concrete topics with a potential to get the MPG prepared for the coming challenges
- this includes infrastructure topics that are essential
- who dares to doubt that unique & persistent identifiers (PIDs) are essential in a research world that moves slowly from a down-load first to a true cyberinfrastructure community?
- not really the most exciting topic

What are the challenges?

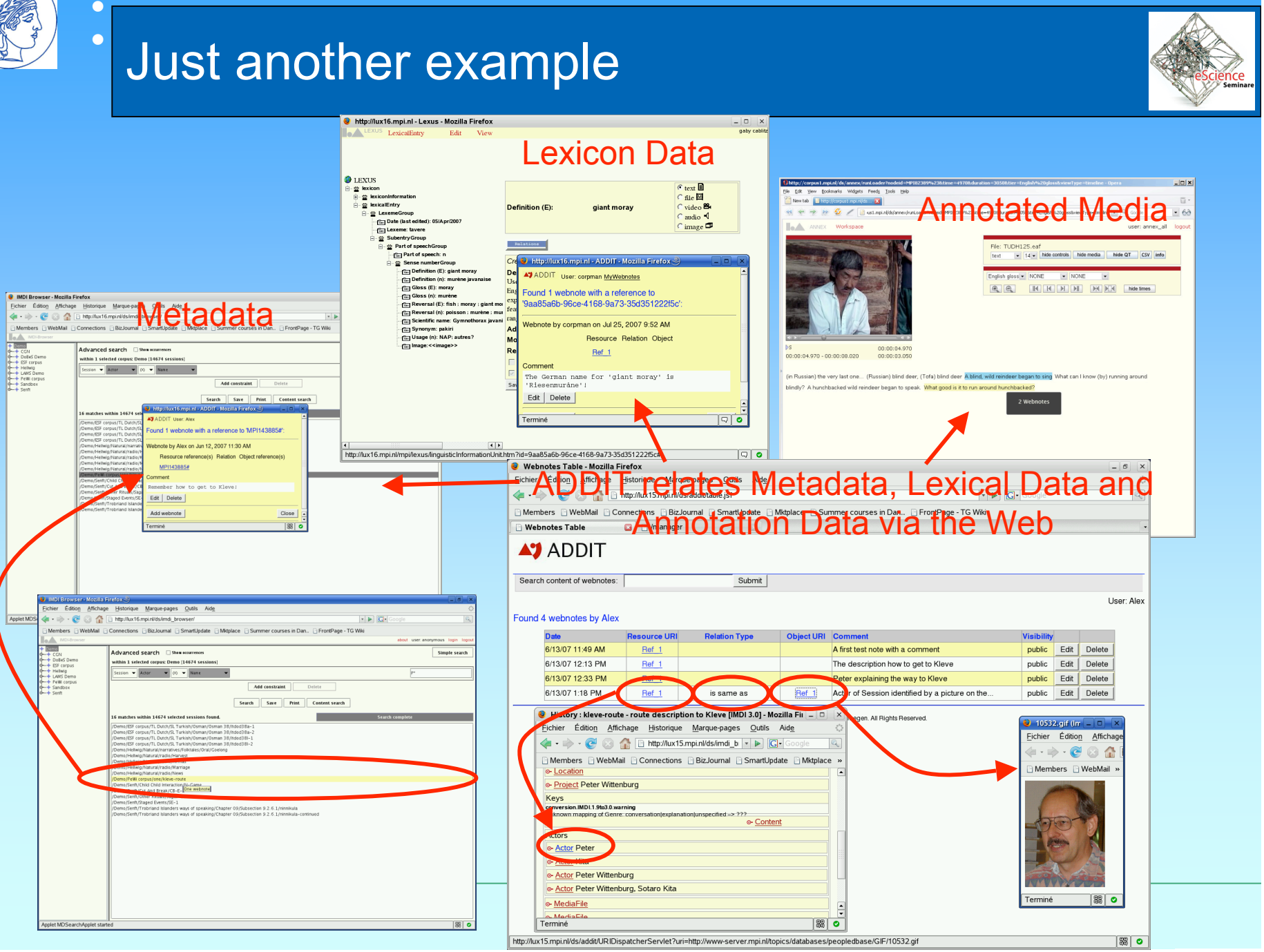
- the primary challenge for research:
  "how can we create a sustainable life given the many increasingly complex and global interdependencies"
- LeveIt:
  "in SSH the main question is whether and how we can maintain a sustainable society – need to update our image of human identity continuously"
- ESFRI: living in the "*century of the complex systems*"
- the fuel for advances in complex model building and testing is primary and secondary data – ever increasing amounts of data
- need to refer to relevant data, need to link between data elements to preserve relevant findings, need to link between documentations to data elements etc

Just a simple example

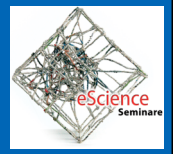
- assume that we have a recording of an extinct language and some annotations that tell us what someone said about medicine etc
- researchers create relations that need to be preserved



Just another example

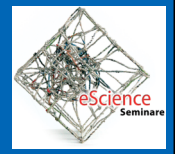


When is it a problem?



- for general use of the Internet there is no serious problem if we end up in a dead link – just use Google or so to find alternatives
- if the research work is carried out in a project no problem either
  - on the fly agreements were possible
  - quick & dirty solutions could be achieved
- even if the research work was done in one institute solutions were possible due to a central control
  (although many cases already where this option failed)
- dead links are an issue if we work crossing boundaries in an anonymous scenario where we need to blindly rely on each other in particular when machines make use of the information
- research infrastructures such as CLARIN all include large communities relying on each other

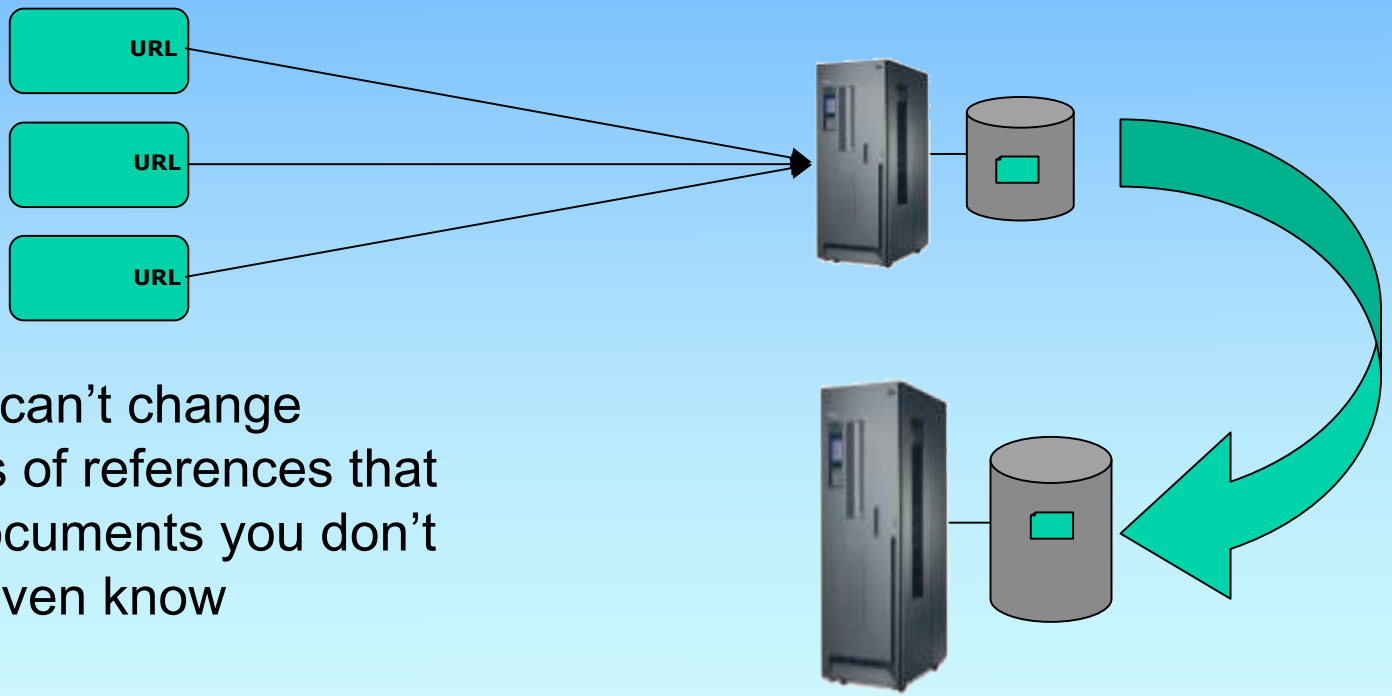
Why are PIDs essential for research?



- if we believe that research is relevant for giving answers to the major challenges of the future then we need to
  - create many important resource relations of different sorts
  - invest expensive researchers time to create these relations
  - preserve these relations to inform our successors
- question1: are our current mechanisms equivalent
- note: the web and the popular URLs were not created to serve the above mentioned intentions
- but since years we are “misusing” it for our purposes and start recognizing some gaps
- question2: can we improve and what is the price?

The problem

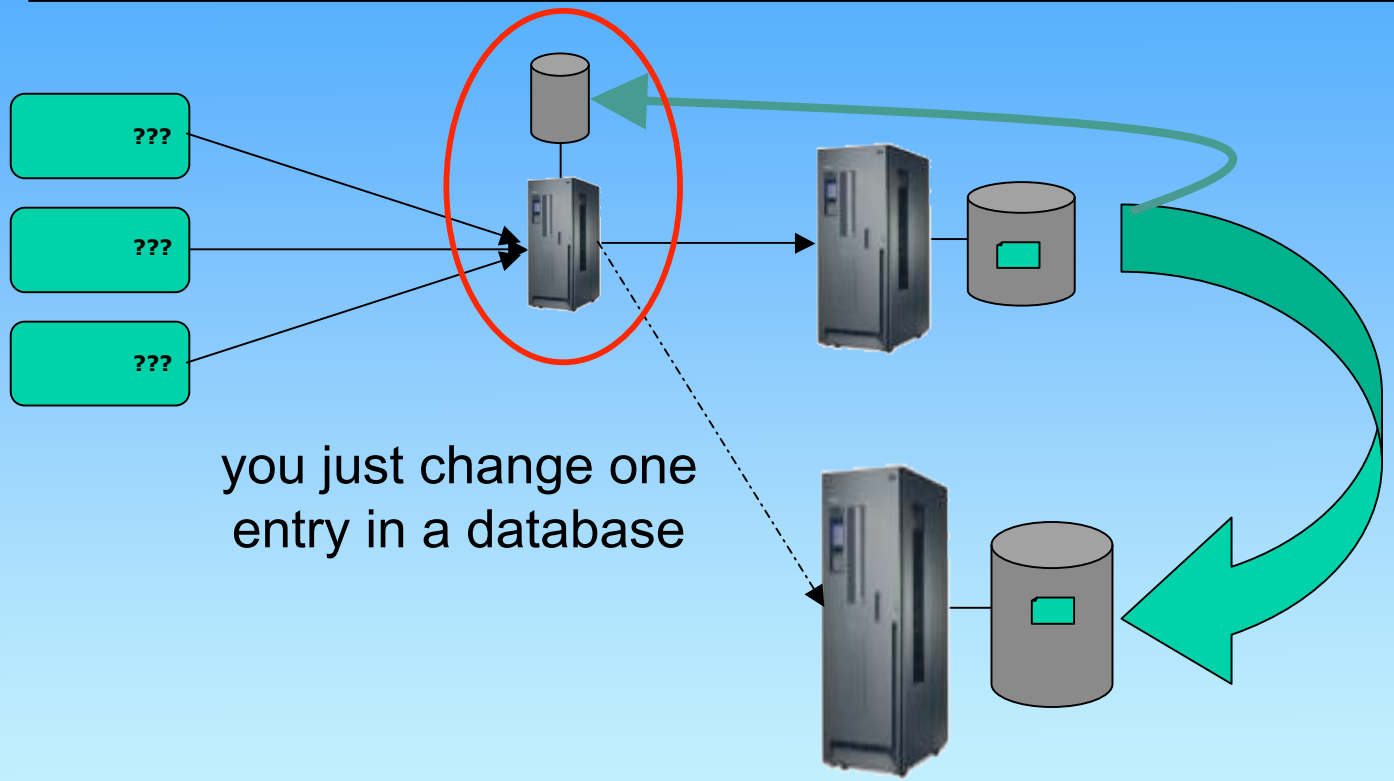
- digital collections are living and subject of continuous migrations
- where to settle the stability required for long-term survival?



you can't change thousands of references that exist in documents you don't even know

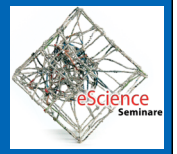
the usual technology migration every 5 years, copying activity or so

Is this a better solution?



- can introduce a new layer of complexity so that the world of references can remain stable despite all changes in the backyard
- price is to maintain this additional complexity
- note: don't introduce a singular point of failure

What are the questions for today?

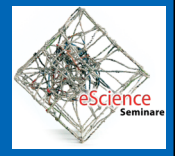


- what are the objects we need to reference?
- do we need an additional layer of complexity?
- which suggestions for PIDs are suitable for research?
  - just numbers vs. semantics included
  - which syntax
  - do we need fragment identifiers
  - do we need to associate additional information
  - which architectural solutions are there
- how complex are the solutions in terms of technology?
- how robust and accepted are solutions?
- what is the business model for institutes with > 500.000 identifiable resources?
- may be even more questions

Programm

- the intention MUST be to work out recommendations for the MPIs
- will need to write a note to the BAR (IT Council)
- so not the last discussion about PIDs within MPG
- therefore excellent program with excellent speakers
- first day
  - views of relevant groups:
    Felix - W3C, Sue Ellen - ISO, Larry – Handle System,
    John – ARK, Norman – DOI, ?? - DNB-Frankfurt
  - panel discussion
- second day
  - contributions from various MPIs
  - summary and conclusions (?)

Let's start



Thanks to our guests for coming and contributing.

Wish us interesting and clarifying discussions.