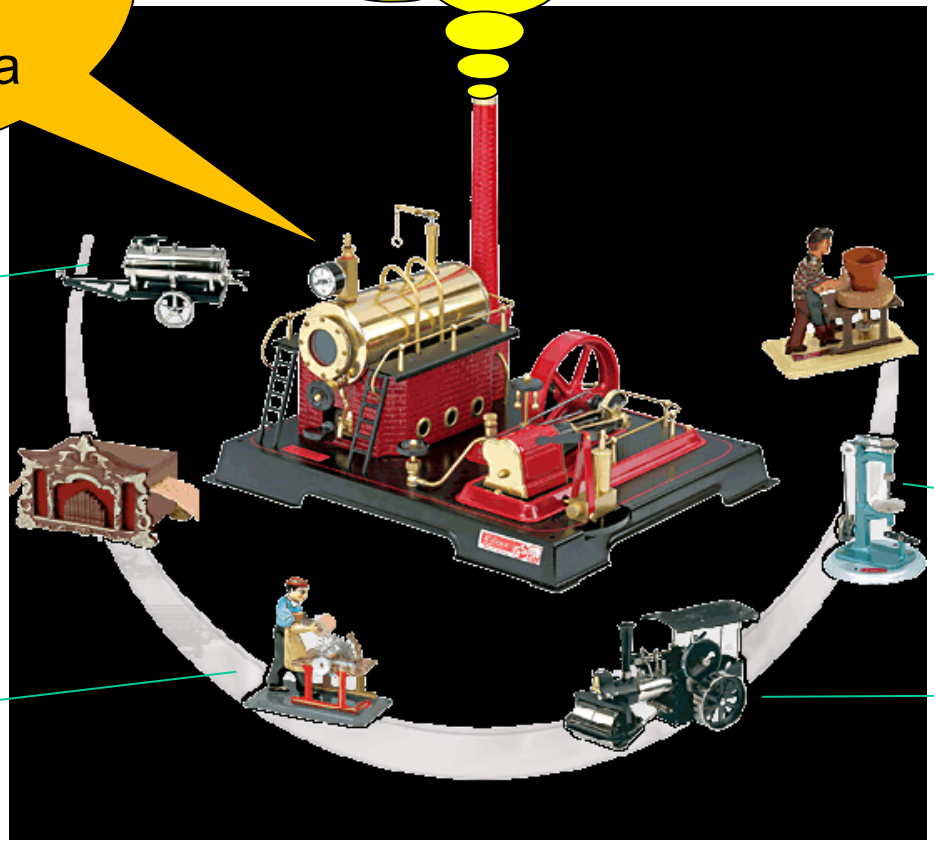
Trust, Costs Issues etc

Peter Wittenburg, MPI for Psycholinguistics

The Research Machine

the running research machine producing and consuming data

publications



biology

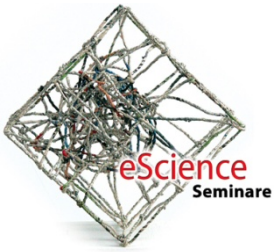
chemistry

medical

humanities

law

physics



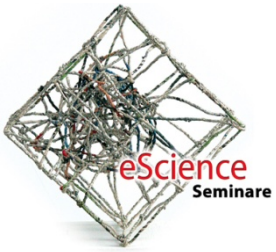
Data ≠ Publications

- publications are different compared to data

- it's the GOLD and key to success in data driven research
- it's not the research result, but one of the main ingredients
- data is dynamic i.e. continuously changing
- the way researchers look at them is not predictable

- publications are the results
- they are "kind of static"

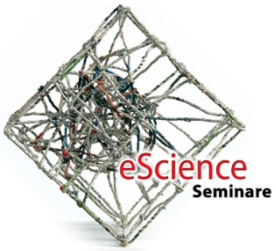
- this all means that the level of trust associated with data is different



Accessibility?

- researchers want fast and convenient access
  - differences what convenience means (workflow, ...)
- therefore still true: first download all stuff on my machine
  or store all my data myself
- no trust in the web due to too many problems and low speed
- of course there is psychology: my data on my machine

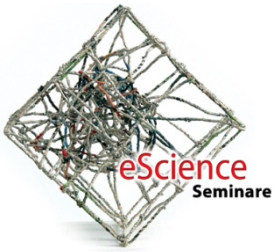
- centres have tendency to introduce burocracy
  - all sorts of explicit agreements (access, availability, etc)
- can we rely on agreements
  (politicians may change rules, companies get bancrupt)



long-term accessibility?

- long-term accessibility has many aspects
  - maintenance of the bitstream
  - interpretability, i.e. format migration to support up-to-date software
  - does data organization survive (metadata, relations)

  - is it online accessible or just for the researcher
  - is access too slow, i.e. practically not accessible
  - is the availability 100%
  - what else?



Is sharing wanted?

- hmmmmm

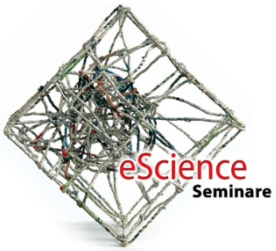
- situation not totally clear
- if it is in the advantage of the researcher sharing is ok
- when does it offer advantages
  - virtual collections of distributed resources
  - cross-disciplinary activities

- is it dangerous
  - www is a world without acknowledgements
  - www is a world where stealing is common practice

Where to store data?

- do we really need repositories??? researcher's ideal is disk

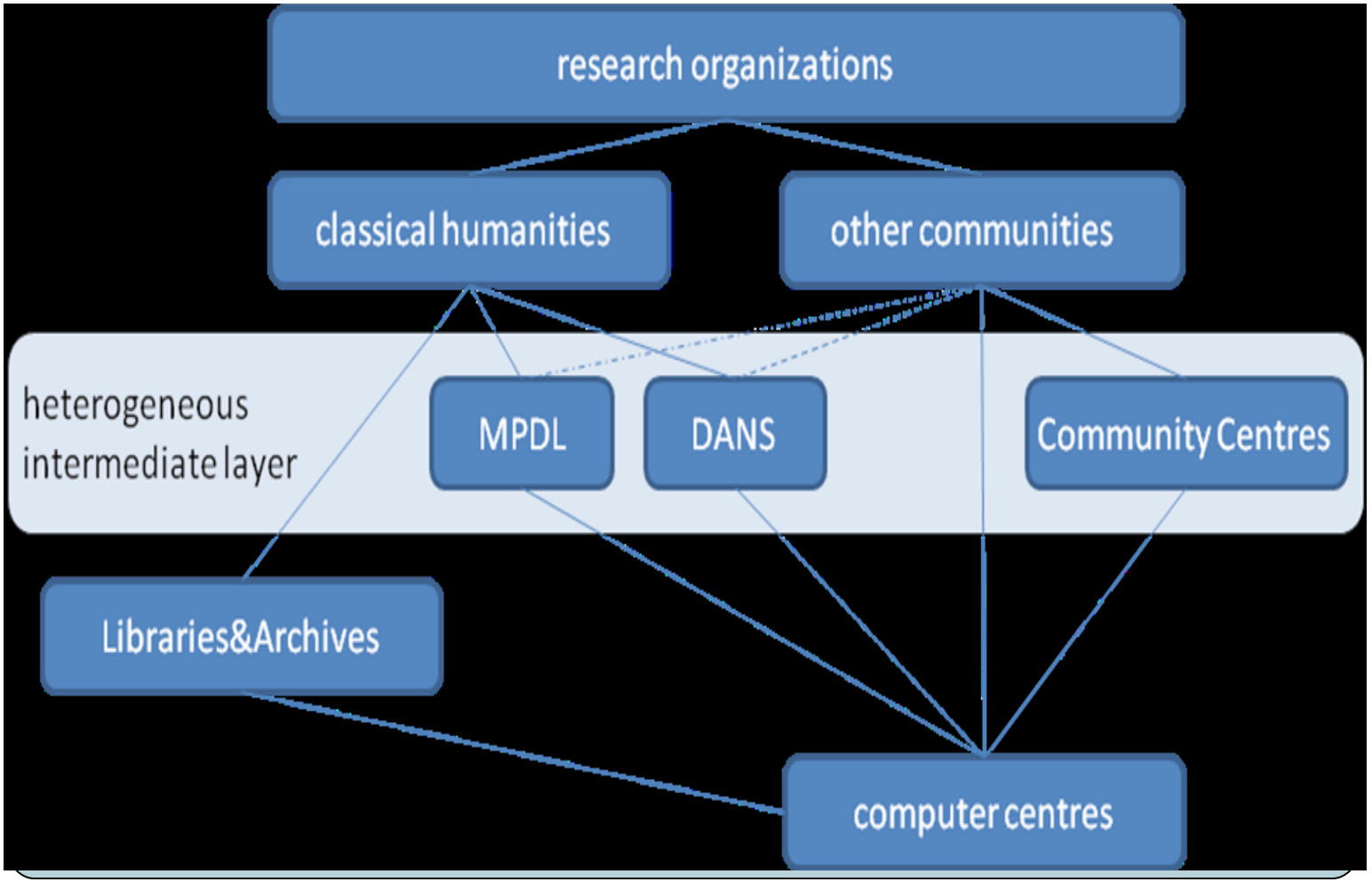
	long-term	accessibility	sharing	trust	costs
private disk	low	high	low	high	?
institutional rep	low	high	moderate	moderate	?
organization rep	high	moderate	high	?	?
community rep	?	moderate	high	?	?
commercial rep	low	?	high	low	?



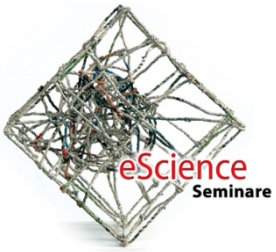
Business models and costs?

- repositories need to have “business models”
  - rules for deposits and access (some forms)
    - researchers hate burocracy
  - rules for maintaining data (replication, migration, etc)
  - rules for rights (look at Google Doc terms)
    - researchers are often very naive
  - service will cost some money
    - do we know the real costs?
    - who will pay?
  - funding schemes are dependent on institute types

Who are the players?



the ama-zoogles



What are the costs?

- machines need to be maintained
  - migration every x years
  - you need people to run the machines
- the “collections” need to be maintained
  - you need people to do it
  - you need software etc for curation and migration
- you have software
  - repository software (there is no good sw for free!!)
  - application software (exceeds the amount of code of the rep system general)

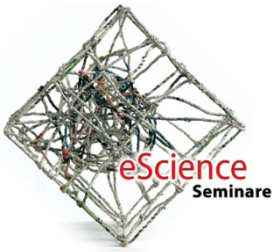
Some numbers - Neil Beagrie

Institutional Repository (e-publications):	Staff	Equipment (capital depreciated over 3 years)
Annual recurrent costs	1 FTE	£1,300 pa

Federated Institutional Repository (data): Annual recurrent costs	Staff	Equipment (capital depreciated over 3 years)
Cambridge	4 FTE	£58,764 pa
KCL	2.5 FTE	£27,546 pa

Some numbers Neil Beagrie

Acquisition and Ingest	Archival Storage and Preservation	Access
c. 42%	c. 23%	c. 35%



MPI Nijmegen example

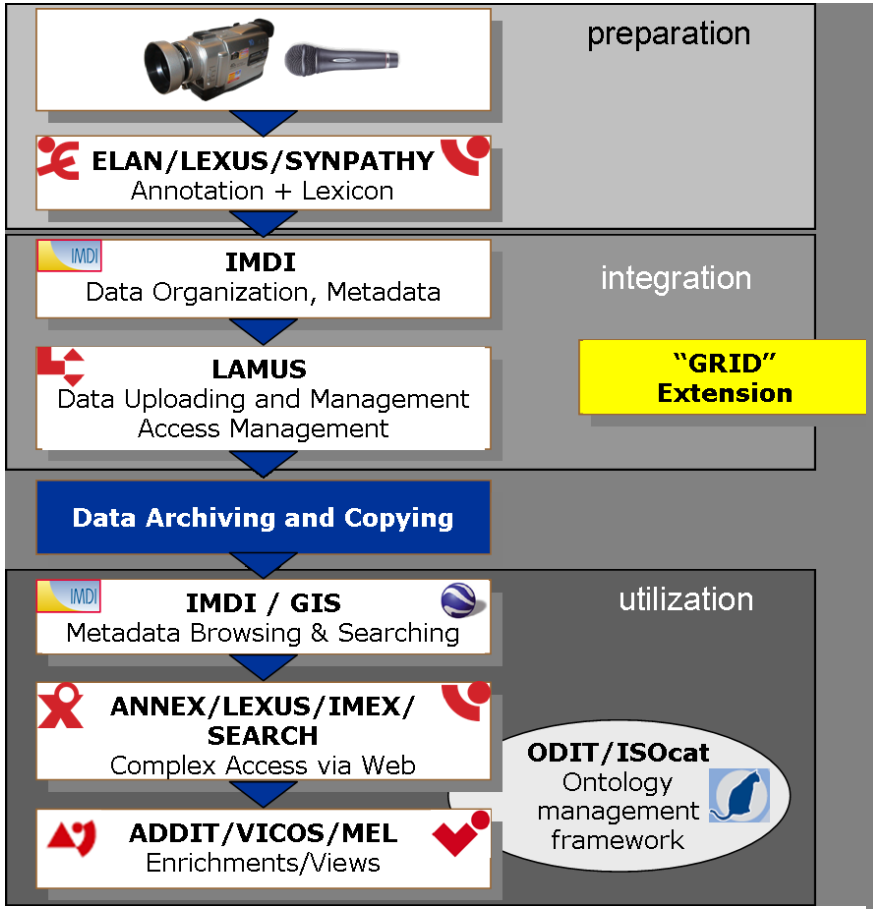
- at MPI start in 2000 as bottom-up process
  - repository system (about 100.000 lines of code)
    incl. metadata, access management etc
    dedicated system tailored to language resources
  - utilization software (about 230.000 lines of code)
    much more heterogeneous
  - rep. system now used by about 13 institutes
  - creation costs of LAT Suite: ~ 1 M€
  - sw maintenance costs for repository system about 60k€/y

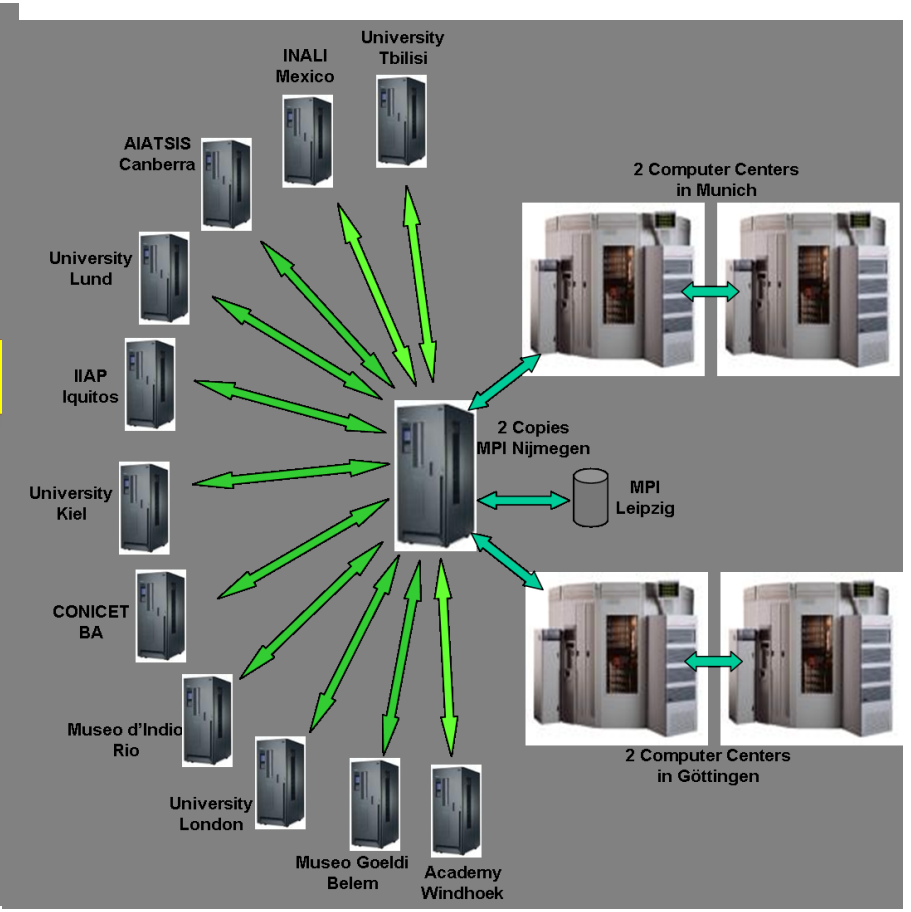
- repository system is core - need maximal independence (MPG, EU)

- what are the costs of other developments and archives?
- costs eScidoc much higher - probably needs a shared model

MPI Nijmegen example

- MPI Nijmegen with complete LAT software suite and replication/migration strategy as an example

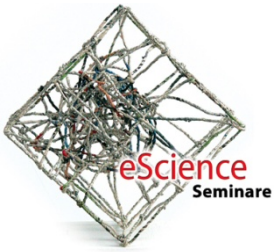




MPI Nijmegen costs

type	k€/y	comment
basic IT infrastructure	80	4-8 years innovation cycle
digitization and workflow	10	new recorders, capturing dev
copies at large computer centers	<5	
system management	60	shared for different activities
archive management	80	advice, curation, consistency
repository software maintenance	60	without new functionality
utilization software maintenance	>120	wide spectrum of tools
building, energy, etc	?	ignored here
total	415	

economy of scale applicable.
(linguistic support, SW development, head etc. not calculated)



End

Thanks for the attention.