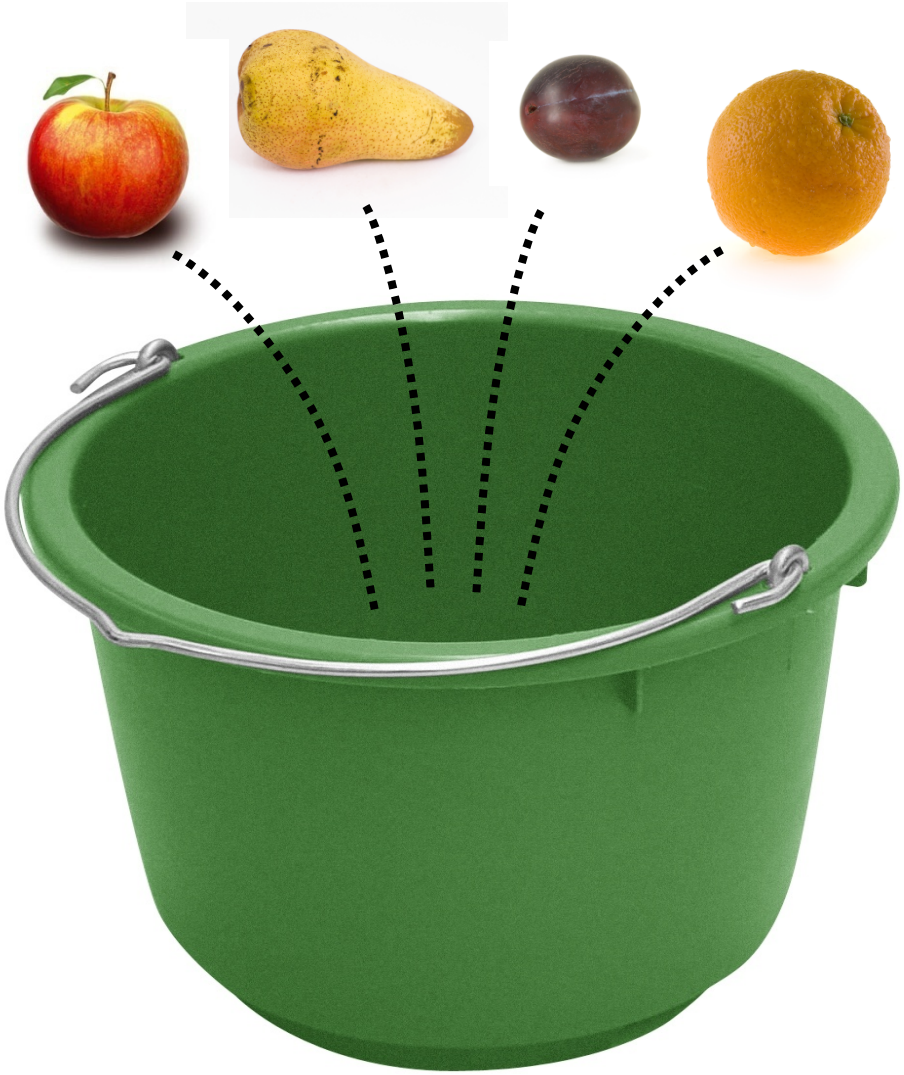
Repository Systems

Peter Wittenburg, MPI for Psycholinguistics

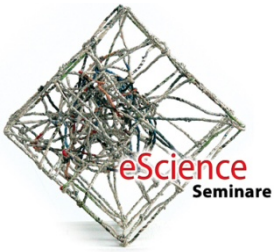
Stefan Heinzel, RZG Garching

What is it all about?



- isn’t it just a pot where to put things in and to get them out again?
- didn’t we do it since years?

- is something changing?
  - type of data
  - amount of data
  - complexity of data
  - sharing of data
  - long-term awareness
  - Internet - security - trust



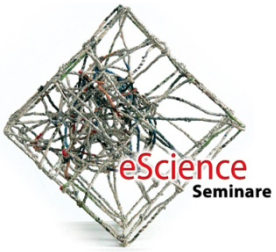
Let's have a closer look I

- put something in AND get the same out again

- do we get indeed the same something - are we sure?
  - did we get the right one?
  - did the thing change?
  - will we get it after 5 years?

- can we change or delete it? - research is dynamic!!!
- what if someone refers to it?
- is it the same when the presentation changes?

- what is the state: in CLARIN 80% cannot give guarantees



Let's have a closer look II

- didn't we do it for years?

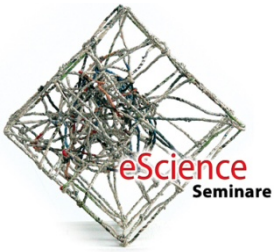
- various flavors
  - give something a file name and a directory path
  - give something a few attributes and put it in rDB which does some encapsulation, compression etc.
  - special systems such as AMOS (Advanced Multi-user Operating System, Friedrich Hertweck, 70ies) to manage and access large amounts of fusion data
  - etc

Let's have a closer look III

- is something changing - type and amount of digital data ?

- publications (well known type - not in focus at this WS)
- from time series to unstructured texts (large variety)
- dynamic collections
  - each individual object can be part of many collections
  - research data is dynamic - continuously new versions

- increasing amount is well-known
  - in genetics for example 1000 genomes: 1 PB/y)
  - a field-linguist studying a language has easily more than 5000 files he is working on
  - how can we find the object and its relatives again

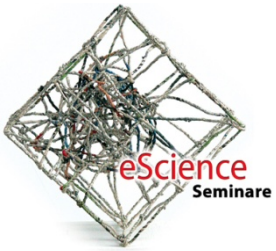


Let's have a closer look IV

- is something changing - complexity of digital data ?

- versions have a close relationship
- extensions have a close relationship
- extractions have a close relationship
- metadata and annotations are closely related
- lots of relationships due to metadata and content dependencies/similarities

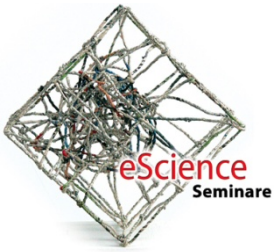
- how to store all these relationships
  - MPI does it by metadata - thus crucial for all aspects



Let's have a closer look V

- is something changing - sharing of digital data and trust?

- people want to share some of their data
- Internet allows sharing across law boundaries and ethical habitudes
- open access is nice - often not practical for primary data
- Internet is anonymous - can we trust users
  - some security is important
  - need to build up trust



Let's have a closer look V

- is something changing - long-term awareness?

- well - increasingly more people believe in the need of preserving research data
- reality is different but ...
  - MPG CC offer archiving service since 2004
  - only few institutes make use of the service

- nevertheless - repository systems can't ignore ltp
  - how long to store - how to do it
  - how to check authenticity despite continous migration
  - how to do replication - physical vs. logical

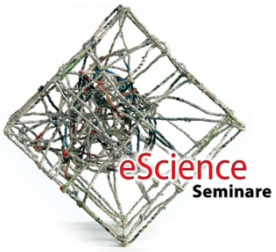
Is the topic seen as relevant?

- JISC:

A digital repository is a managed, persistent way of making research, learning and teaching content with continuing value discoverable and accessible. Repositories can be subject or institutional in their focus. Putting content into an institutional repository enables staff and institutions to manage and preserve it, and therefore derive maximum value from it. A repository can support research, learning, and administrative processes. They are commonly used for open access research outputs.

- Forrester Research:

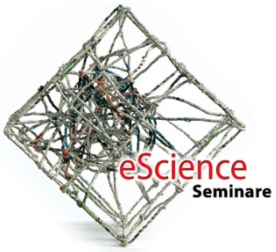
“Knowledge workers spend 40% of their time trying to find information and 70% of that time is spent recreating information that cannot be found. A digital repository offering refined categorisation and search tools that help locate information quickly provides quantifiable savings in terms of time and resources.”



Is the topic seen as relevant?

- ESFRI (European Strategy Forum on Research Infrastructures)
  - WG established criteria for digital repositories
    - Availability: data and metadata must be available
    - Permanency: preservation, management and curation
    - Quality: policy for data quality
    - Rights of Use: clear statements about accessibility
    - Interoperability: support of open standards

- e-IRG (Infrastructure Reflection Group)
  - WG established document about
    - guidelines for metadata
    - guidelines for quality check and assurance
    - guidelines for interoperability



Is the topic seen as relevant?

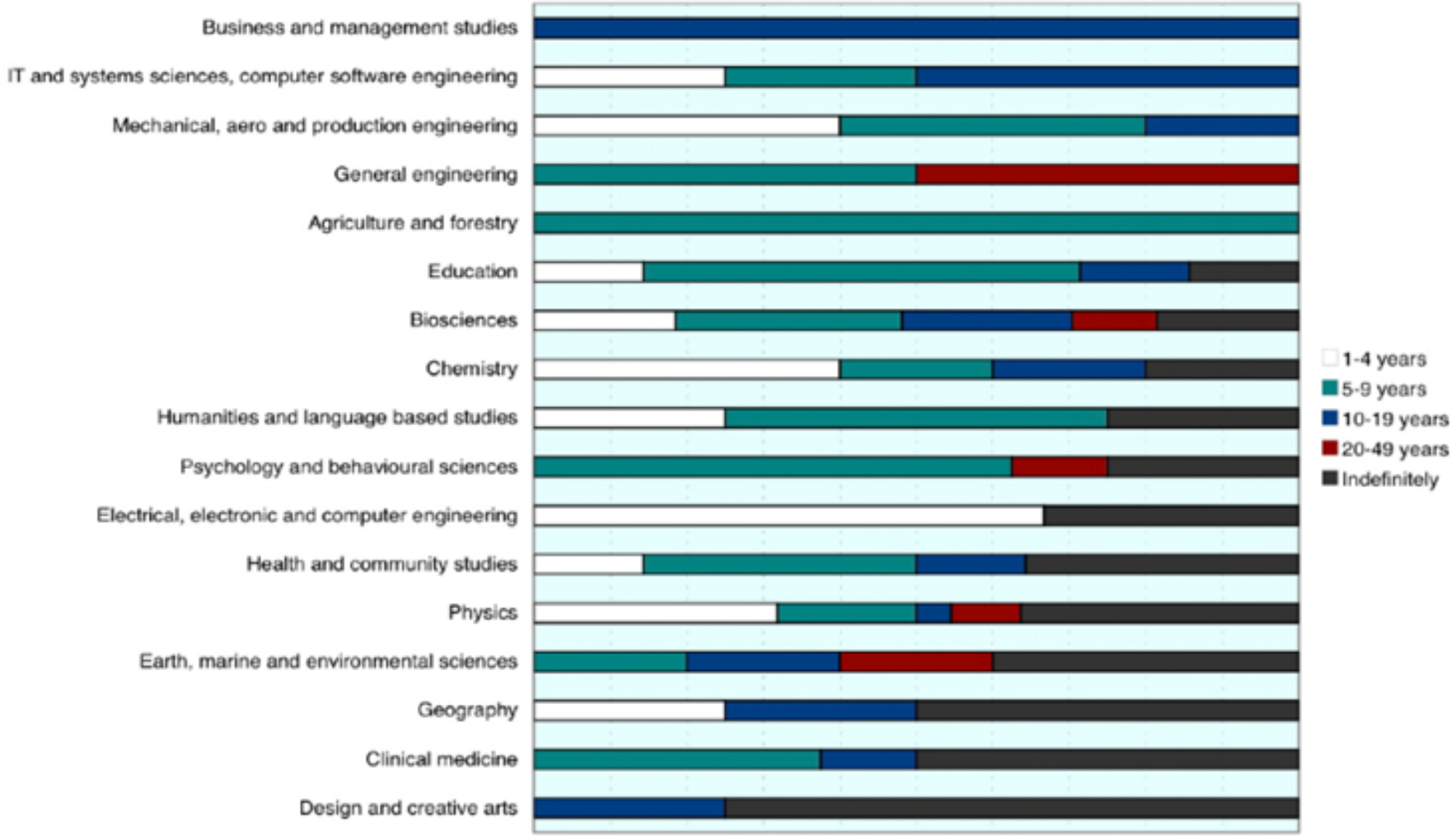
- DELOS (Expert Network on Digital Libraries)
  - Digital Library Manifesto (DL ≠ repository)
  - come up with complete framework description
  - developed a reference model and an abstract API
- a number of software systems:
  - FEDORA: flexible object model as basis for repository system
  - eScidoc: a realization of a FEDORA bases system by MPDL/FIZ
  - D-SPACE: ready made system mainly for publications
  - ePrints: similar
  - iRODS: strong system with grid/federation functionality
  - LAMUS: a tailored system by MPI
  - MS: has an offer (as few other companies do)
  - etc etc

Is the topic seen as relevant?

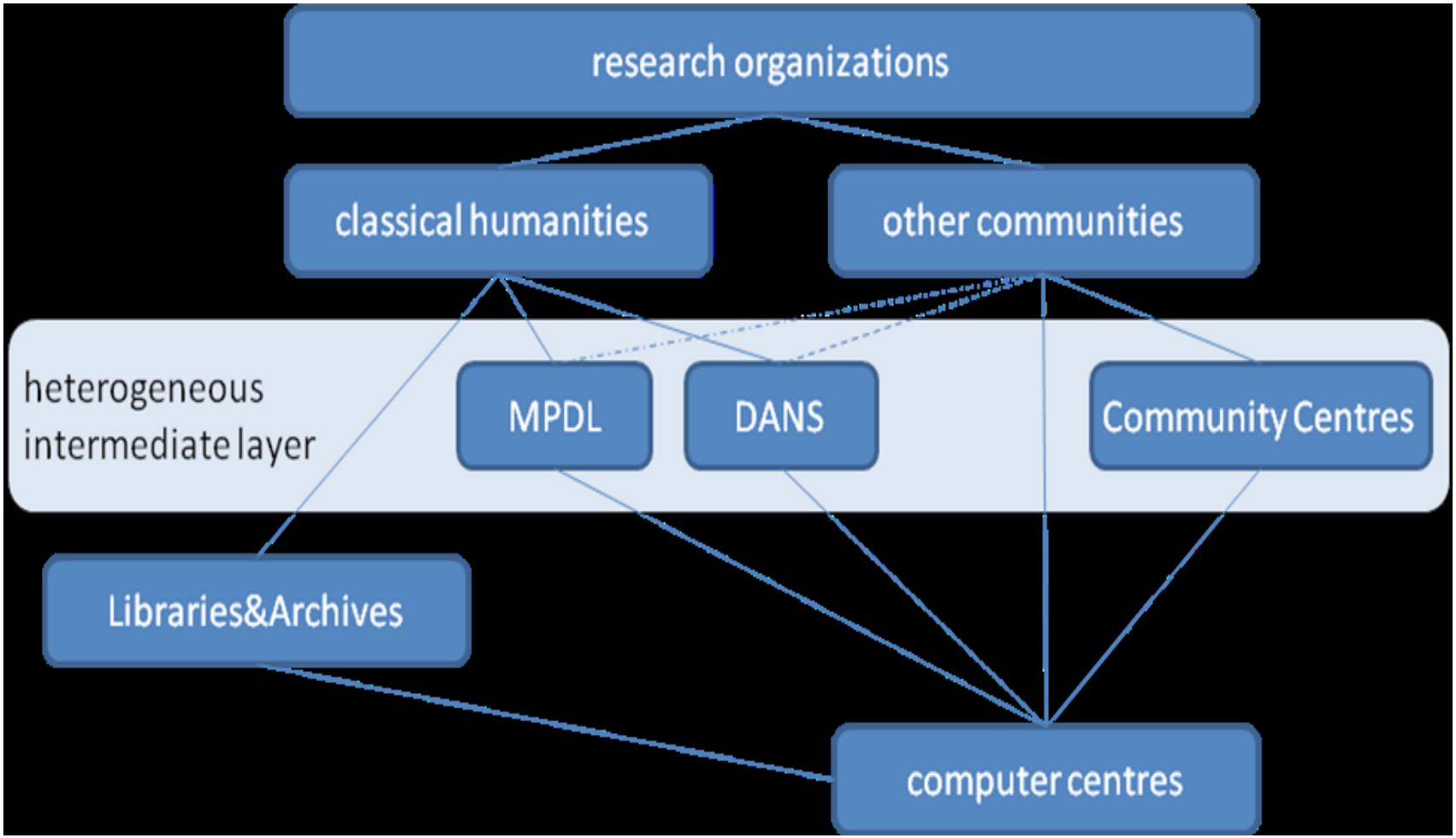
- DRIVER survey:
  - 80% of repositories have publications
  - 10% also have primary/secondary research data
- CLARIN:

  25 well-known centres want to become node
  but only few have a proper repository system!
- conclusions:
  - general discussion very much library oriented
  - awareness changing right now

Is the topic seen as relevant?

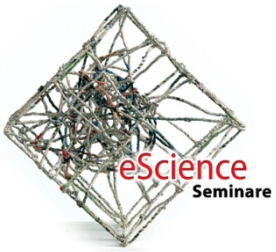


Who are the players?



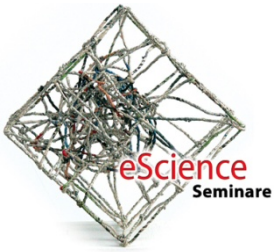
What was mentioned?

	CLARIN	LIFE-WATCH	DIFRE	ELIXIR	INCF	Climate	Space	STFC	EDIFIS
repository, long term archiving, authenticity	👍	👍	👍	👍	👍	👍	👍	👍	👍
discovery, access, data mining, virtual integration, curation	👍	👍	👍	👍	👍		👍	👍	👍
high performance computing, common processing layer	👍			👍	👍	👍	👍		
data distribution, data federation, data grid	👍		👍	👍	👍	👍	👍	👍	👍
high availability, high reliability	👍	👍					👍	👍	
SSO, S ID, trust AAI	👍	👍	👍			👍	👍	👍	
PID support	👍	👍	👍			👍	👍	👍	
(component) metadata	👍	👍	👍			👍		👍	
SOA, web services, workflow, interoperability	👍	👍	👍	👍	👍	👍			
format and semantic interoperability, standards	👍	👍	👍	👍	👍	👍			
network of domain nodes	👍		👍	👍		👍	👍		



Typical Requirements for Research Reps

- support of high security and access permission mechanisms
- subject of quality assessments
- management and retrieval support by hq metadata (provenance etc)
- support of building arbitrary collections
- support of live archiving principles (changes, enrichments, relations)
- support for format migration
- each object needs to be identified by a PID
- authenticity checks need to ensure object identity
- support for replication at logical level with PID pointing to copies
- researches need to rely on availability and persistence
- federation support
- no resource encapsulation
- support of APIs for program interaction
- cost efficient (rep menagament, code maintenance, etc)
- etc

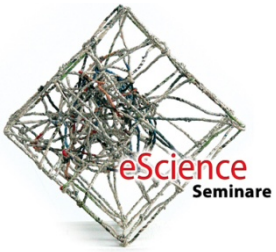


Program Day 1

Time	Topic	Speaker
10.00	Introduction + Requirements	
10.45	Necessity of Repositories and Roles	Wouter Spek
11.30	ESFRI & e-IRG Requirements	Dany Vandromme
12.30	Lunch	
13.30	Data Models and Abstraction Layers	Carlo Meghini
14.30	Repository Solution from MS	Savas Parastatidis
15.30	Coffee	
16.00	Solution/Plans at Meteorology	Frank Tussaint
16.30	Solution/Plans at Psycholinguistics	Daan Broeder/ Mariano Gardellini
17.00	Solution/Plans at Cognitive and Brain Science	Roberto Cozatl
17.30	Solution /Plans at Biophysical Chemistry	Holger Bartels
18.00	End	
19.00	Dinner	

Program Day 2

Time	Topic	Speaker
9.00	eScidoc Architecture	Malte Dreyer
9.45	Repositories/ Federation in iRODS	Adil Hasan
10.30	coffee	
11.00	Requirements for a Repository System	Ulrich Degenhardt
11.45	Long-term Requirements and Quality Assessment	Henk Harmsen
12.15	Lunch	
13.15	Trust, Accessibility, Costs etc	Peter Wittenburg
13.45		
14.30	Wrap Up, Discussion	Andreas Gros, Frank Toussaint
18.00	End	



End

Goal is to right a report and provide a guideline for MPIs.

Let’s start then.