# e-IRG Panel

# Data Infrastructure - Grids - HPC
# DAITF

Peter Wittenburg – Max Planck Society, Germany

# Where do I talk about?

- data-oriented researchers create/look for suitable data (small - big collections)
- data is stored in a variety of centers (community data centers - common data centers)
    - EUDAT: ideally copies of all ”registered objects” will be in ”registered ” data centers
    - data centers requirements: persistent, certified, robust, service oriented, etc.
- researchers then want to execute smart operations and workflows on the stored data
- thus what research communities need are frameworks that allow users
    - to virtually integrate and access distributed & interdisciplinary collections (CDI)
        - based on visibility, identity, registered syntax and semantics
    - to execute automated workflows on these collections in the data centers
    - to quickly and dynamically deploy services close to the data
- ”true HPC” is different - although storage of the data in common data centers
- data centers may apply any technology (cloud, grid, ...)
    - must be transparent to users and cheap (cost participation)

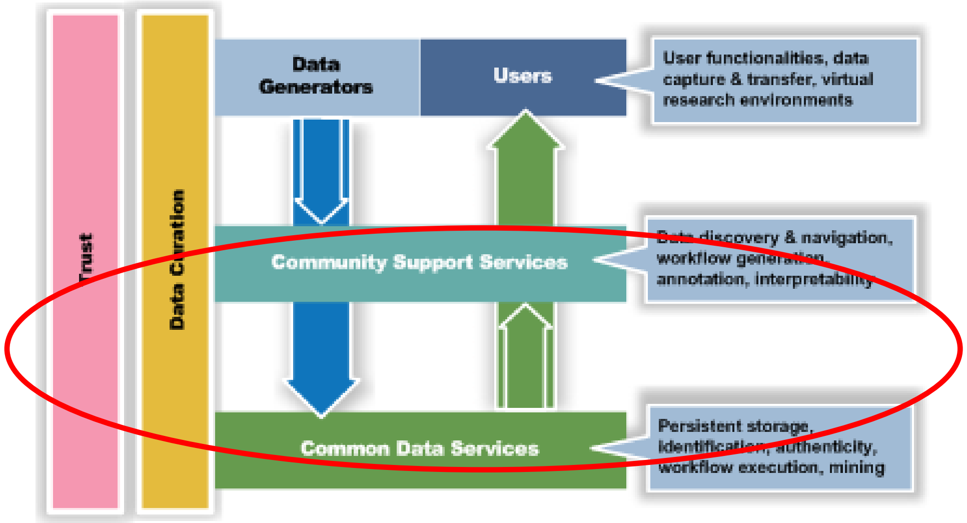
# to the questions ...

2. Data Infrastructures start from the needs of data oriented research - thus it is a pillar
    - certainly different user communities with different priorities
    - HPC = highly parallelized ”big” jobs;  Grid community = throughput
3. Heterogeneity of Data Infrastructures does not serve all user wishes
    - except for ”islands” we lack even basic elements of an open data object domain
    - trust & integration mechanisms are lacking - not speaking about interoperability
4. Committed research communities and providers need change of cullture
5. Integration data-grid-HPC process must primarily be bottom-up driven
6. Future infrastructures MUST include a commitment from RO
    - either from the beginning or after 3 years (EUDAT example)
    - need for European sustainability & equalization funds are relevant
7. Sustainability of living research collections needs to be based on RO commitments
    - bit-stream preservation vs curation

# What keeps us busy?

## Collaborative Data Infrastructure

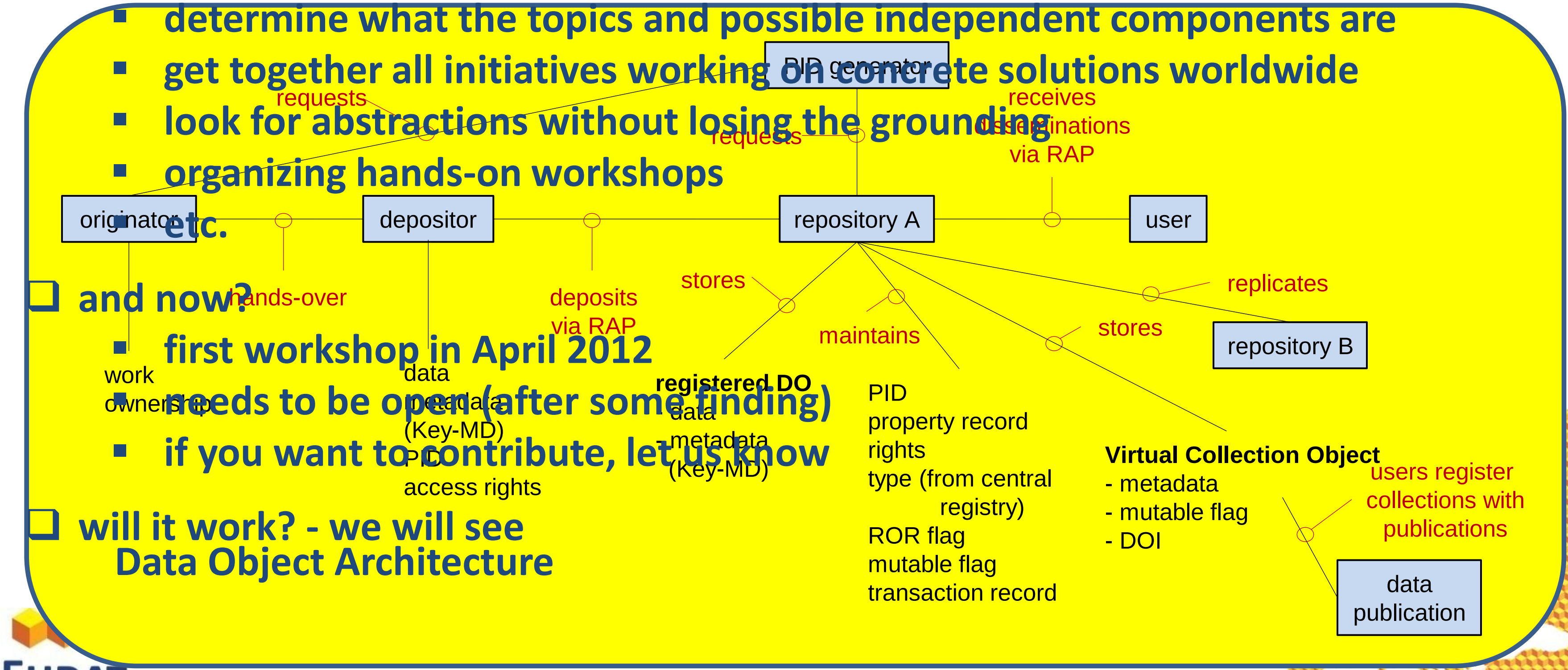
Source: HLEG report, p. 31


Trust
Data Curation
Data Generators
Users
User functionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation, workflow generation, annotation, interpretability
Common Data Services
Persistent storage, identification, authenticity, workflow execution, mining

- ❑ virtual collection and workflow building requires integration & interoperability
  - ▪ this is a "hard" problem
  - ▪ yet we did not even agree on a domain of clearly registered digital objects (some (CNRI, ITU, EPIC, DOI) are working on a worldwide registration service)
- ❑ look at the e-IRG Data Management Task Force report
  - ▪ chapter 3: 2 levels of interoperability (resource + semantic level)
  - ▪ world is certainly more complex dependent on view (infrastructure levels, metadata at object and content level, etc)
- ❑ there are many groups now working on CDI - all facing similar challenges
  - ▪ need a "common data object architecture" (Bob Kahn)
  - ▪ need abstractions to realize CDI integration layer
  - ▪ need syntax/API wrappers and flexible semantic mapping frameworks
  - ▪ etc

# Is DAITF a way?

- ❑ DAITF = Data Access & Interoperability Task Force
  - obviously IETF as good example for a grass-roots based approach
  - governance etc just as much as is needed
  - but data world is more complex/heterogeneous than network world or?
- ❑ basic DAITF ideas
  - find a common language (example: data object architecture)
  - determine what the topics and possible independent components are
  - get together all initiatives working on concrete solutions worldwide
  - look for abstractions without losing the grounding
  - organizing hands-on workshops
  - etc.
- ❑ and now?
  - first workshop in April 2012
  - needs to be open (after some finding)
  - if you want to contribute, let us know
- ❑ will it work? - we will see



Data Object Architecture