“The EUDAT Research Agenda”

David Corney (STFC, UK) WP7 Task leader

EUDAT USER FORUM, 7-8 MARCH 2012

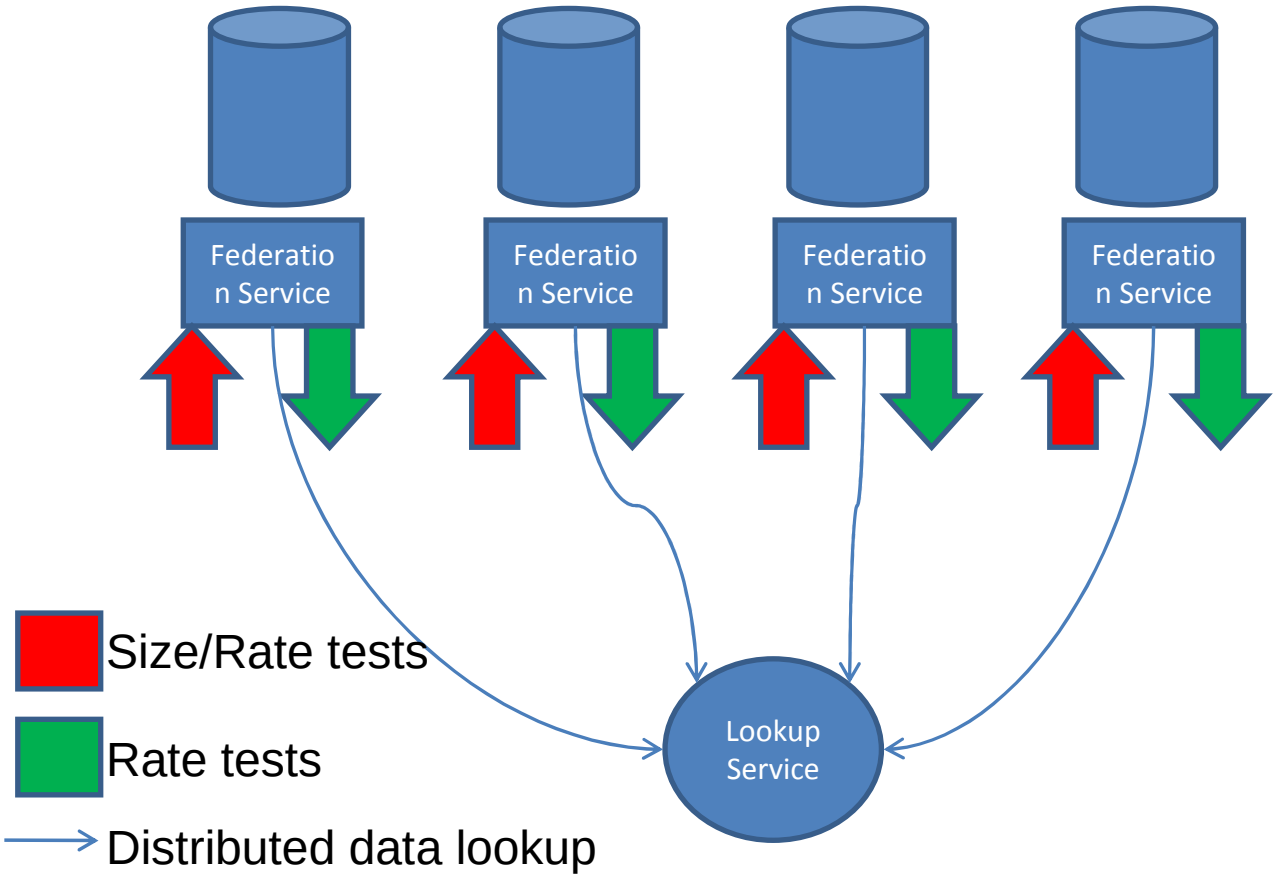
EUDAT

Date:

High level objectives

- “Addressing Scalability and Preservation”
- Objectives:
  - Enhancing the Core CDI architecture
    - Scaling federation across existing archives
    - Federated data curation and long term preservation
  - Resolving known scalability issues
  - Developing a scaleable workflow engine
- 3 – 5 year perspective

Scaling federation across existing archives



Scalable test prototype :

- Link (multiple) existing archives (<= 10^9 files per archive)
- Link (multiple) existing metadata catalogues
- Minimise disruption to existing infrastructures

Scaling federation across existing archives

- Test infrastructure going into place (STFC, PSNC, DKRZ, RZG, CERN)
- Test plan agreed – various technologies:
  - Federation technologies: IRODS, Dcache, Xroot
  - Transfer prorocols: Htpc, ftp, gridftp, http, bbcp, rsync, rbudp, xroot
- Seperate scalable metadata solutions from archive federation
  - Currently being explored via Metadata Task force, incl WP7

Data curation & long term preservation in a federated environment

- Example test scenario:
  - A data manager at a community center receives a collection of data enhanced with complex metadata defining data hierarchies in the collection. The data collection itself is a coherent set of files, however the complex hierarchies defined by the metadata can also be stored in a database during ingestion

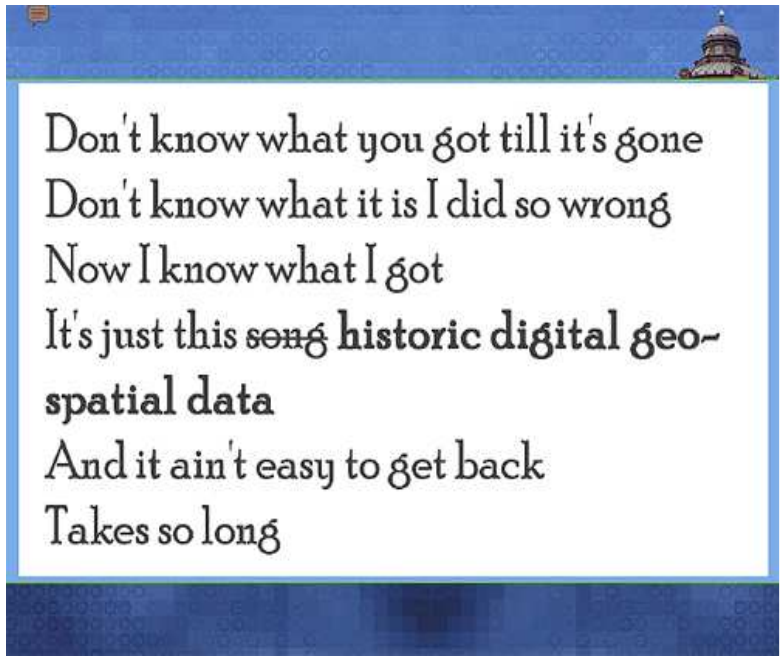
- Example policies:
  - Convert subset files to different file formats at primary data centre
  - Replicate *n* copies of the data in *m* different data centres and enable user access to certain data sets
  - Enable complex hierarchy definitions to be stored in *m* different data centres

Data curation & long term preservation in a federated environment

- Extend policy complexity:
  - Replicate access permission information (on application level – not file level)
  - Manage the associated PID records - Information about the new location of all replicas should be appended to the PID recor
- Federate
  - Further extend policy complexity beyond initial test centres (CLARIN/MPI-PL and ENES/MPI-M&DKRZ) by introducing new centres (EPOS, VPH, ….)
  - Coordinating with scalable federation work undertaken by 7.1.1

2r7

3



Diapositiva 6

dr7 Not clear what this means? Please can you clarify
corney; 24/02/2012

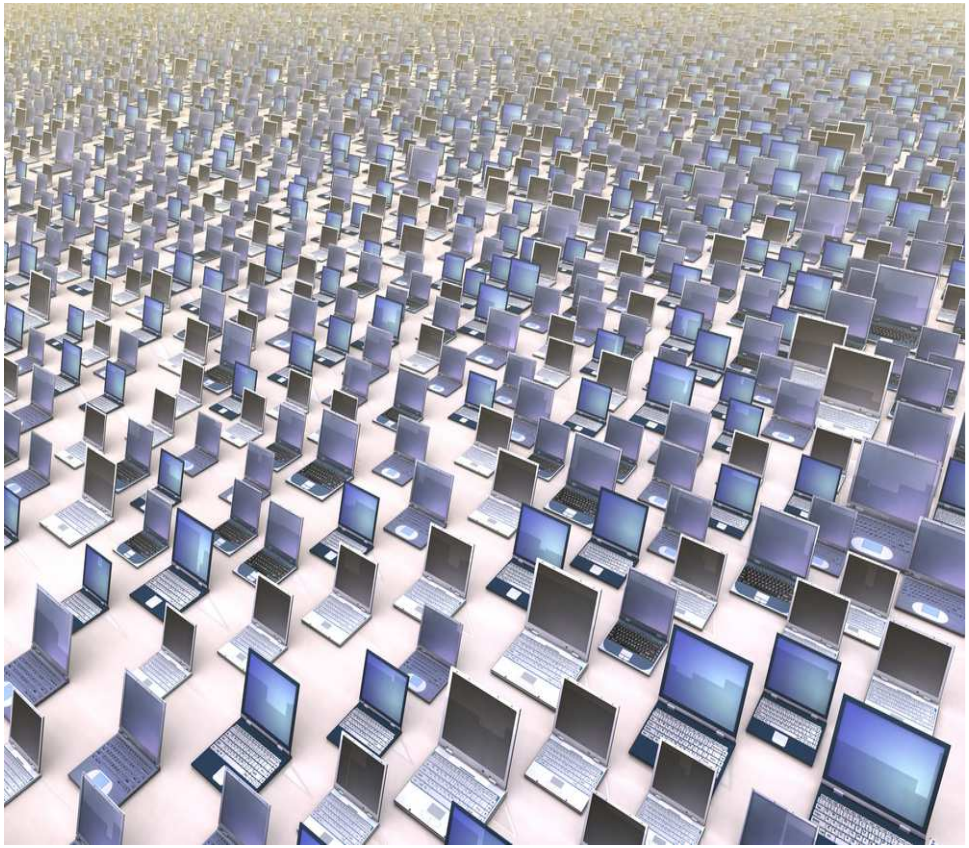
2 We are talking about the replication of authorization information for the application layer. This layer operates on top of the data and therefore we are not talking about permissions on the filesystem.

Wether this item is something for WP7.1.2 or the AAI taskforce is another discussion.
Willem Elbers; 27/02/2012

3 The initial development of this will be done within the TFs / WPs. Within WP7.1.2 we should focus on evaluation this solution and the integration of more communities into this solution.
Willem Elbers; 27/02/2012

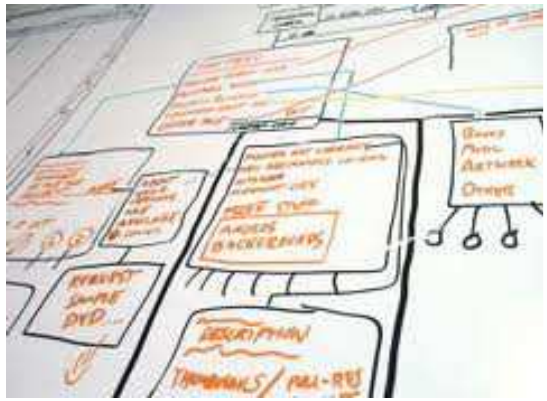
Scalability issues

- Interfaces to data archives:
  - Optimise and simplify interfaces to scalable archives
  - Develop a road map for:
    - Common archive interfaces across multi user communities
    - A new archive interface standard
- Scalable data access:
  - Optimise data access technologies across federated archives
    - Replication, data caching algorithms
- Provide user transparent access to optimal copy of data



Developing a scaleable workflow engine

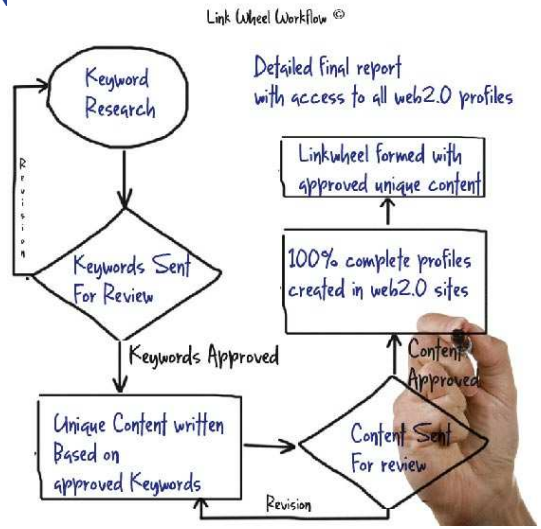
- Avoid download TBs of unecessary data
- Overcome the wide variety of work flow systems - a barrier to shared analysis across communities



- Deliver a generic execution framework:
  - Compatible with existing work flow systems
  - Compatible with existing tools and data formats
  - Map community created workflows into a generic form via a common interface
  - To locate remote data sets of interest (link to7.1.1)
  - Integrated with web services to access and process remote data sets
  - Work flow executes close to data sources or across distributed data resources
  - Support efficient data streaming to manage large volumes (rather than memory based or file based processing)
  - Returns analysis results (only) to user

Developing a scaleable workflow engine

- Currently
  - Evaluating technologies (to cope with data scale)
    - Benchmark tests defined (based on ENES & CLARIN requirements)
    - Initial test framework designed and deployed
    - Initial test data sets identified (EKUT
- *Planned:*
  - *Build generic work flow execution framework*
  - *Integrate with:*
    - *Earth System Grid data Servers (work across federated data sets)*
    - *Climate Web services – (manage scalable data volumes)*
    - *Language data technologies (extend to encorporate Weblicht)*




```
jWorkflow.order(garlic
          .andThen(white
          .andThen(wonto
          .andThen(cooki
          .andThen(noAnd
          .andThen(noAnd
          .andThen(noAnd
          .andThen(noAnd
```

Thank you.

WP7 - Deliverables

Month	Deliverable description	Deliverable Id	Task
12, 24, 36	Towards a globally scalable archive federation technology	D7.1.1, D7.1.2, D7.1.3	7.1.1
18, 36	Managing data curation and long term preservation in a federated environment	D7.2.1, D7.2.2	7.1.2
18	Evaluation of existing interfaces to data archives, roadmap for evolution towards a standard	D7.3	7.2
36	Recommendations for techniques and policies to enable optimized data access	D7.4	7.2
24, 36	Technology adaption and development framework	D7.5.1, D7.5.2	7.3

WP7 Milestones

Month	Milestone description	Id
M8, M22	Re-assessment of priorities	MS22, S27
M12	Data exploration technology experiments and benchmarking	MS23
M12	Checkpoint that work on federated archives can begin	MS24
M18	Generic Execution Framework design	MS25
M24	Checkpoint that construction of prototype federated archive can begin	MS26

Aim, position and perspective