

The need for sustainable e- infrastructure

Presentation in NEERI09
conference

Kimmo Koski

CSC – IT Center for Science

October 1st, 2009

CSC – Tieteen tietotekniikan keskus Oy
CSC – IT Center for Science Ltd.

Why e-infrastructure is needed?



- New Research Infrastructures (RI) are being built: 44 in ESFRI list
- Existing RIs are developing: wild estimation 100 – 150 RI with different size
- All need e-infrastructure (computing, data, networks, applications, competence, ...)
- All need the services to be sustainable with guaranteed availability for tens of years
- Worst case scenario: 150 – 200 RI working with incompatible ICT silos
- Target to have a well balanced ***ICT Ecosystem***

Global Virtual Research Communities

Research
Community-1

Human
interaction

Workspace

Labs

Scientific Data

Computing,
Grid

Network

Research
Community-2

Human
interaction

Workspace

Labs

Scientific Data

Computing,
Grid

Network

Research
Community-3

Human
interaction

Workspace

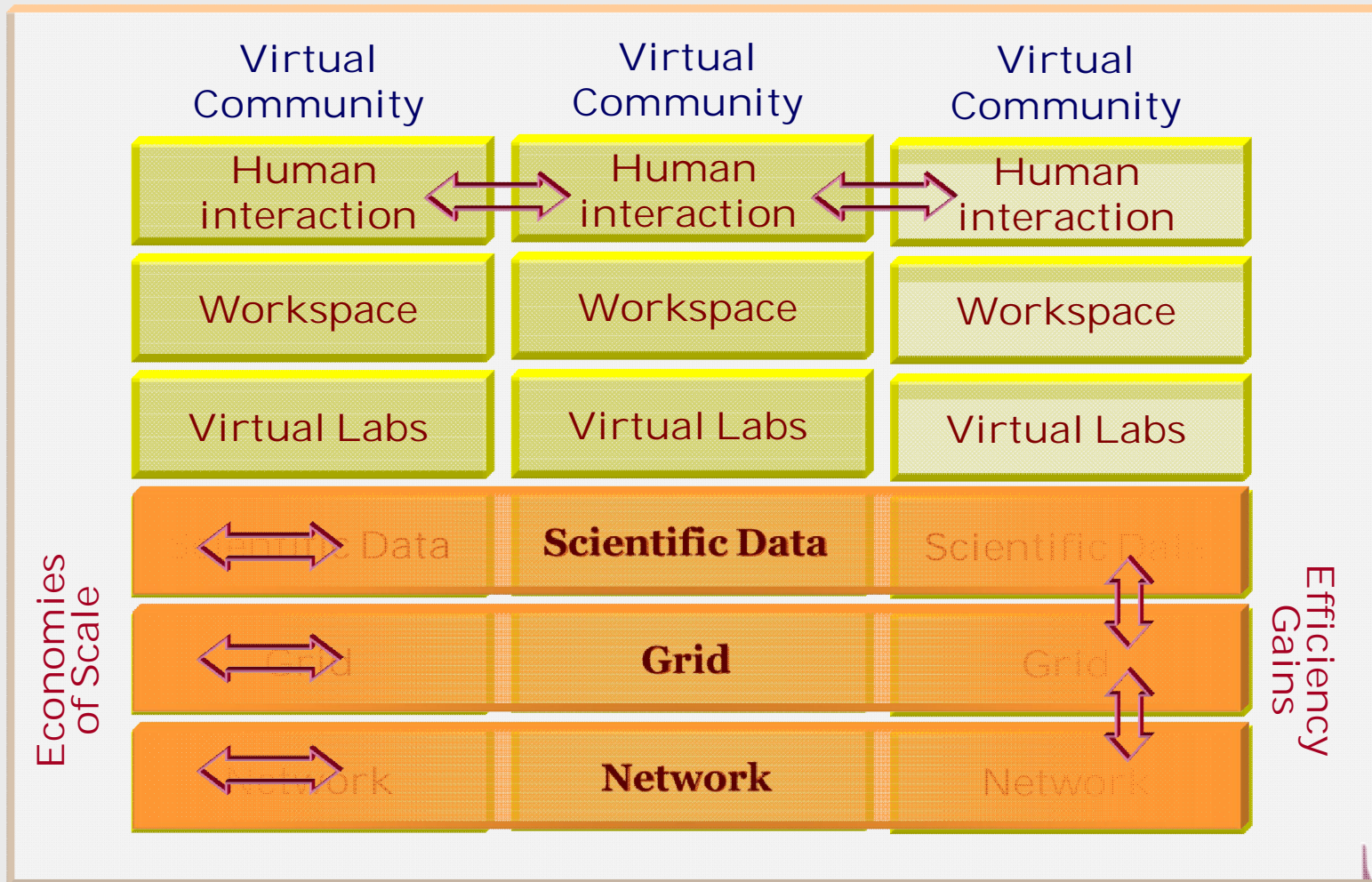
Labs

Scientific Data

Computing,
Grid

Network

Global Virtual Research Communities



Data and information explosion

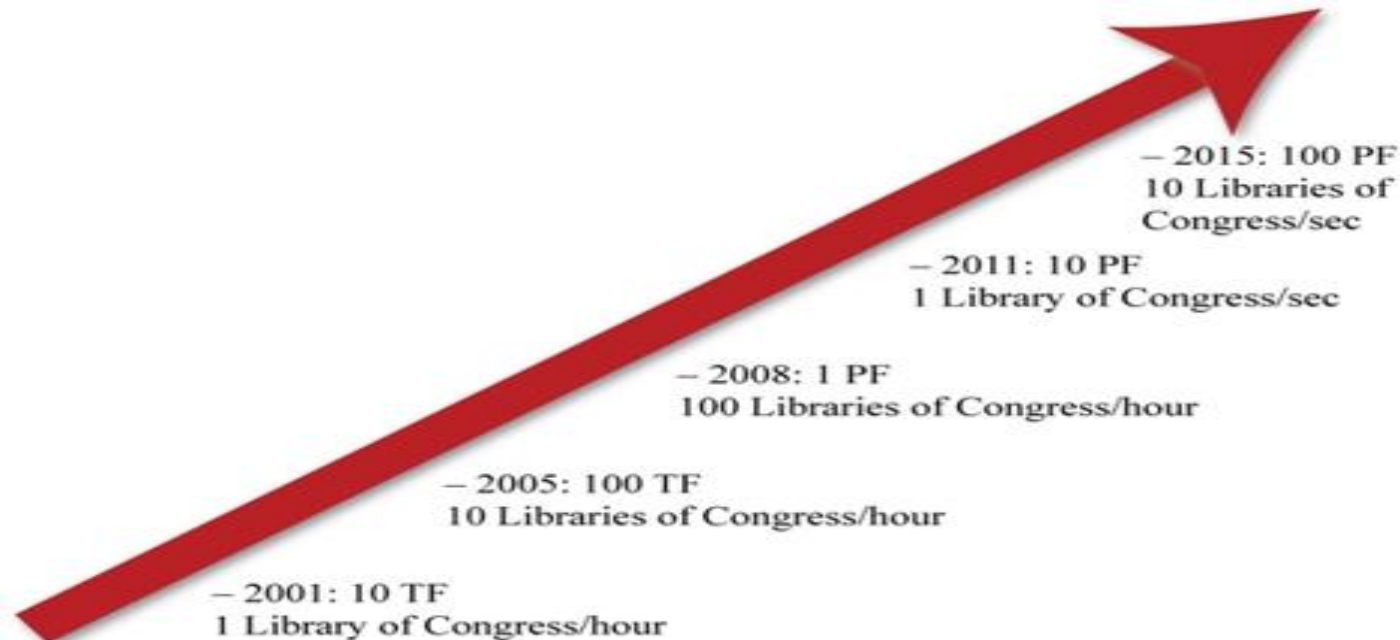
Petascale computing produces exascale data

1 Gigabyte (1GB)
= 1000MB
CD album

1 Terabyte (1TB)
= 1000GB
Word yearly
Book production

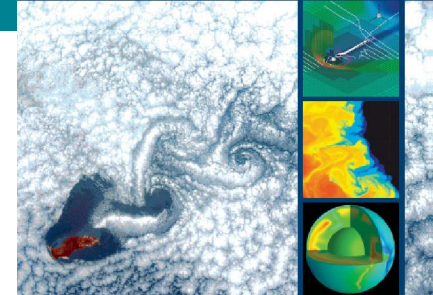
1 Petabyte (1PB)
= 1000TB
One LHC-
experiment
yearly data
production

1 Exabyte (1EB)
= 1000 PB
World yearly
information
production





Unprecedented volumes of data:



Phenomenal growth of data in all various dimensions

Unsolved questions in

Scale

Heterogeneity

Stewardship

Curation

Long-Term Access and Storage



New generation of tools for data discovery, integration, visualization, analysis and preservation



New standards, open protocols and interfaces



New scientific opportunities



New careers: data managers, data scientists, data authors...

Future Aim:

To have scientific digital data routinely deposited in well-documented form, regularly and easily consulted and analyzed, and openly accessible while suitably protected and reliably preserved.

Needs:

- Coherent organizational framework
- Flexible tiered technical architecture:
 - Standard open protocols and interfaces
 - Flexible user access, analysis and visualization of data
 - Address issues of authentication, authorization, security
 - Supports workflows



How to keep the data for ever?

Long term storage

Long-term Storage is the storage of data for a period of time which surpasses the typical life-cycle of the hardware and software components employed, with the *goal of keeping the data forever*.

This typical life-cycle is:

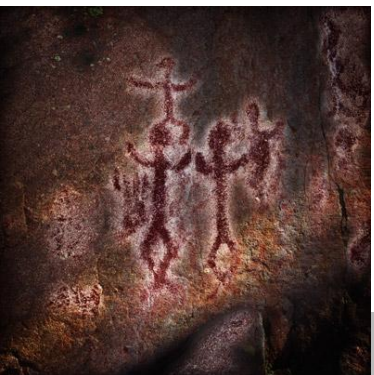
- archive hardware (tape and disk systems): about 5 years
- HSM systems (software and hardware): about 10 years
- filesystems with HSM functionality: about 20 years
- metadata servers (databases): about 20 years
- document formats (.tif, .pdf ...): ??? years

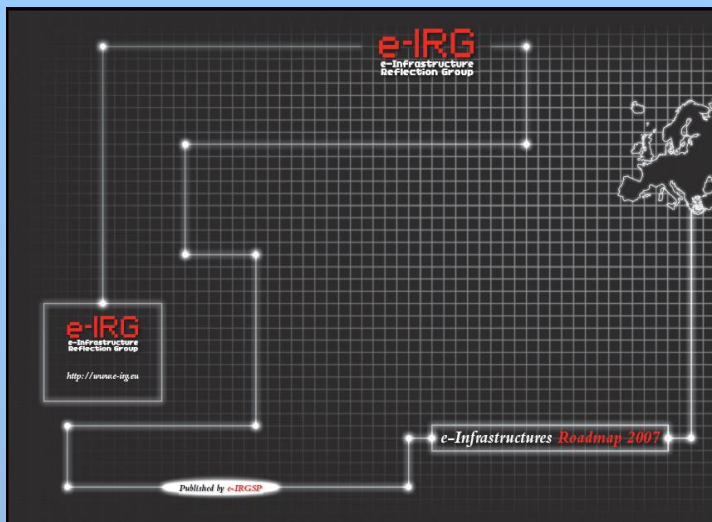
Issue 1: Getting the data to the media

Issue 2: The media

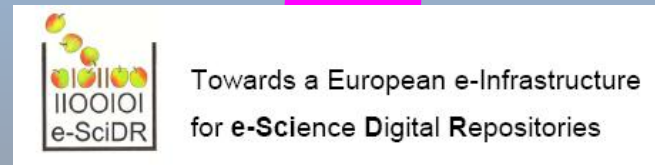
Issue 3: The format

Issue 4: Using the data

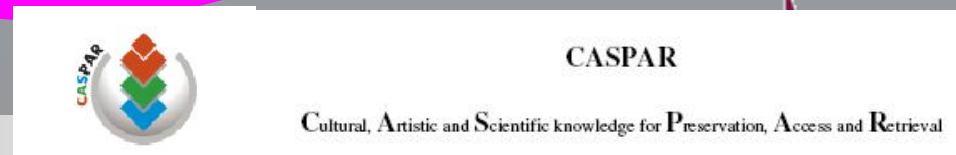




The EU landscape



planets





**COUNCIL OF
THE EUROPEAN UNION**



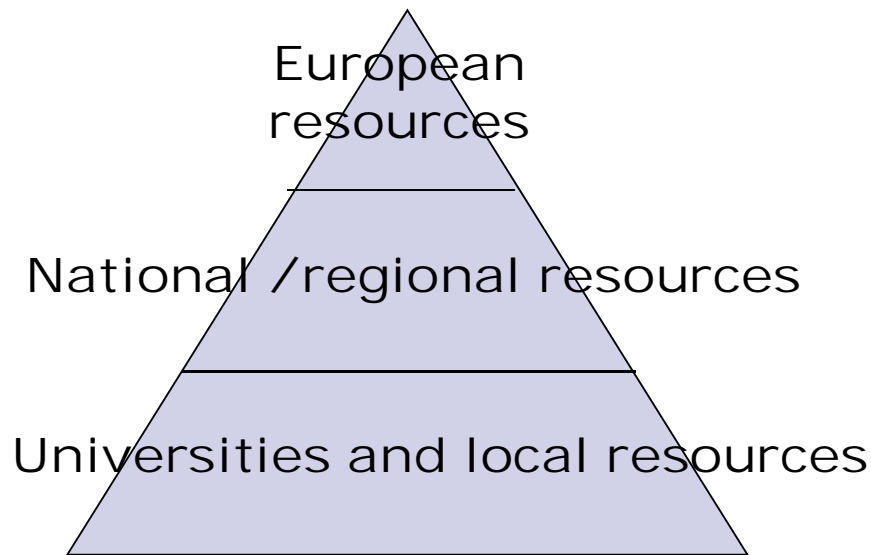
Council Conclusions on scientific information in the digital age: access, dissemination and preservation

*2832nd COMPETITIVENESS (Internal market, Industry and Research) Council meeting
Brussels, 22 and 23 November 2007*

- **reinforce national strategies and structures for access to and preservation and dissemination of scientific information, tackling organisational, legal, technical and financial issues;**
- enhance the co-ordination between Member States, large research institutions and funding bodies on access, preservation and dissemination policies and practices;
- maximise access for researchers and students to scientific publications, in particular by improving public procurement practices in relation to scientific information; this could include exchanging information on these practices and increasing the transparency of the contractual terms of "big deals", and exploring the possibilities for funding bodies, research institutions and scientific publishers from different Member States to work together in order to achieve economies of scale and efficient use of public funds by demand aggregation;
- ensure the long term preservation of scientific information - including publications and data - and pay due attention to scientific information in national information preservation strategies;



Some challenges in building an Ecosystem



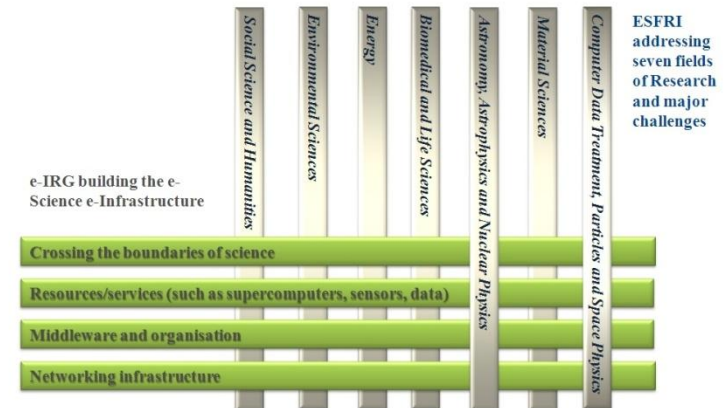
1. How to guarantee access to the top for selected groups?
2. How to ensure there are competent users which can use the high end resources?
3. How to involve all countries who can contribute?
4. How to develop competence in home ground?
5. How to boost collaboration between research and e-infrastructure providers?
6. What are the principles of resource exchange (in-kind)?

Some Key Issues



- Sustainability
 - EGEE and DEISA are projects with an end
 - PRACE and EGI are targeted to be sustainable with no definitive end
- ESFRI and e-IRG
 - How do the research side and infrastructure side work together?
 - Two-directional input requested
- Requirement for horizontal services
 - Let's not create disciplinary IT silos
 - Synergy required for cost efficiency and excellence
- ICT infrastructure is essential for research
 - The role of computational science is growing
- Renewal and competence
 - Will Europe run out of competent people?
 - Will training and education programs react fast enough?

Roadmap to an ESFRI e-Infrastructure eco-system



Policy and strategy work



- HET: HPC in Europe Taskforce
<http://www.hpcineuropetaskforce.eu/>



- e-IRG: e-Infrastructure Reflection Group
<http://www.e-irg.org/>



- ESFRI: European Strategy Forum on Research Infrastructures
<http://www.cordis.lu/esfri/>



- ERA Expert Group on Research Infrastructures

EU infrastructure projects

GEANT



Number of data infrastructure projects



Some focus areas for future



- Energy efficient ICT
- From flop/s to research results
- Data challenges
- Application development
- Competence development for multidisciplinary skills
- Collaboration across disciplines
- Starting from user needs, not from technology

Strategy for European Data Infrastructure, white paper



- Published on Friday afternoon in NEERI
- Drafted by PARADE (PARtnership for Accessing Data in Europe) consortium
- Target to strengthen the European collaboration on data
- Focus on user communities and their requirements driving the service development

Conclusions: Towards a sustainable e-infrastructure



- From projects to long term commitment
 - PRACE, EGI, ESFRI-projects, ... what is for data?
- Based on sustainable elements
 - National funding in a major role
- Horizontal ICT services in a key role
 - We can't afford to reinvent the wheel 150 times
- Providers of e-infrastructure and services succeed only if approaching the technology from user community perspective
 - Need to build trust between the stakeholders