d4science

5° EGEE User Forum
12-15 April 2010
Uppsala (Sweden)

Pro's and Con's of Cloud Computing for Scientific Data – the (on-going) experience of D4Science

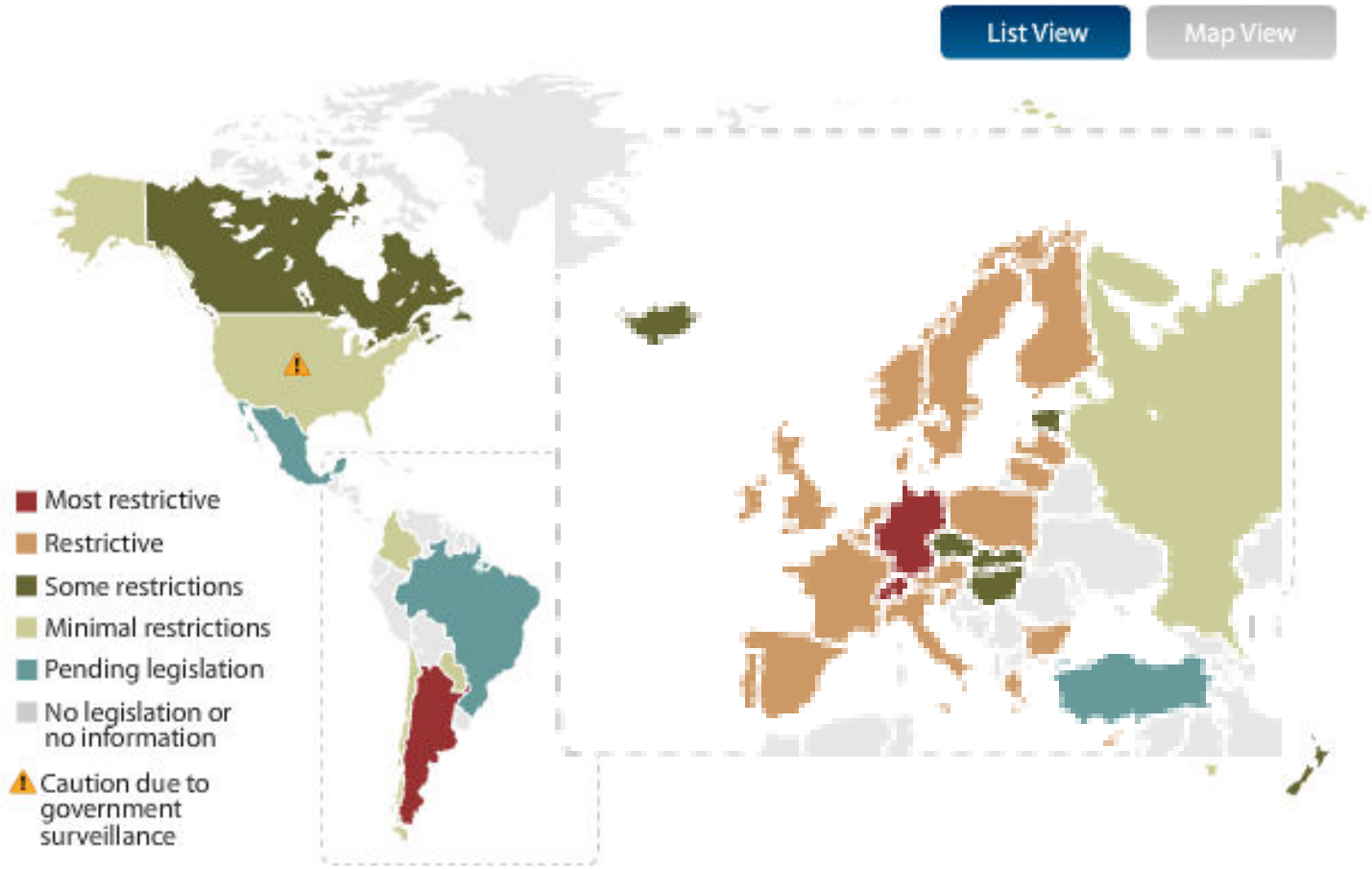
Andrea Manieri
Engineering Ingegneria Informatica s.p.a.

- Cloud Computing: mith or reality?
- gCube evolution meet Cloud technologies
- Scientific Data infrastructures and Cloud
- Towards sustainable Scientific Data Infrastructures

What analysts say

- N.Carr - IT doesn't matters, 2003
  - IT as a commodity - Not marketing, business!
- J.Gray – Distributed Computing Economy, 2003
  - Data near the compution – a economy thread-off analysis
- N.Carr – from EDISON to GOOGLE, rewiring the world, 2007
  - The world wide computer and the implication on economy and society
- Advices and check-lists
  - Garnter, IDC, Forrester-research
  - Security, License schema, ROI

Interactive Data Protection Heat Map



Source: US Department of Commerce and country specific legislation

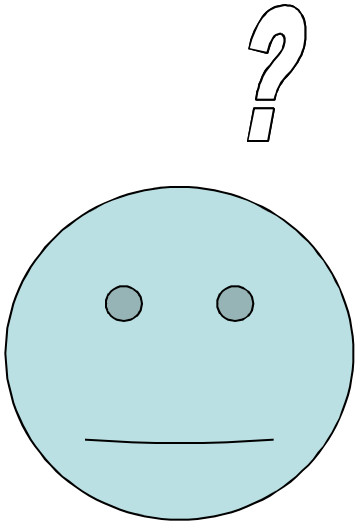
Source: Forrester Research, Inc.

Note: This interactive map provides information on national data protection laws that have either been enacted or are currently under consideration

What experts say

- Old technology with new behaviors
  - The illusion of infinite computing resources available on demand,
  - The elimination of an up-front commitment by Cloud users
  - The ability to pay for use of computing resources on a short-term basis as needed and release them as needed, (*Berkeley Univ. Above the clouds, Jan 2009)*
- Clouds at the baseline of the Future Internet
  - Provided the following gaps are covered: Manageability and Self*; Data management, Security and privacy; federation and interoperability; virtualisation Elasticity and Adaptability; APIs, programming models and remote control (*The future of clouds – report, Jan 2010*)
- Lights and shadows
  - Economic benefits are evident
  - To be careful with security, privacy, and SLA (*ENISA Security and Cloud WG Nov 2009)*

What vendors say



What Community does

OCCI wg within OGF

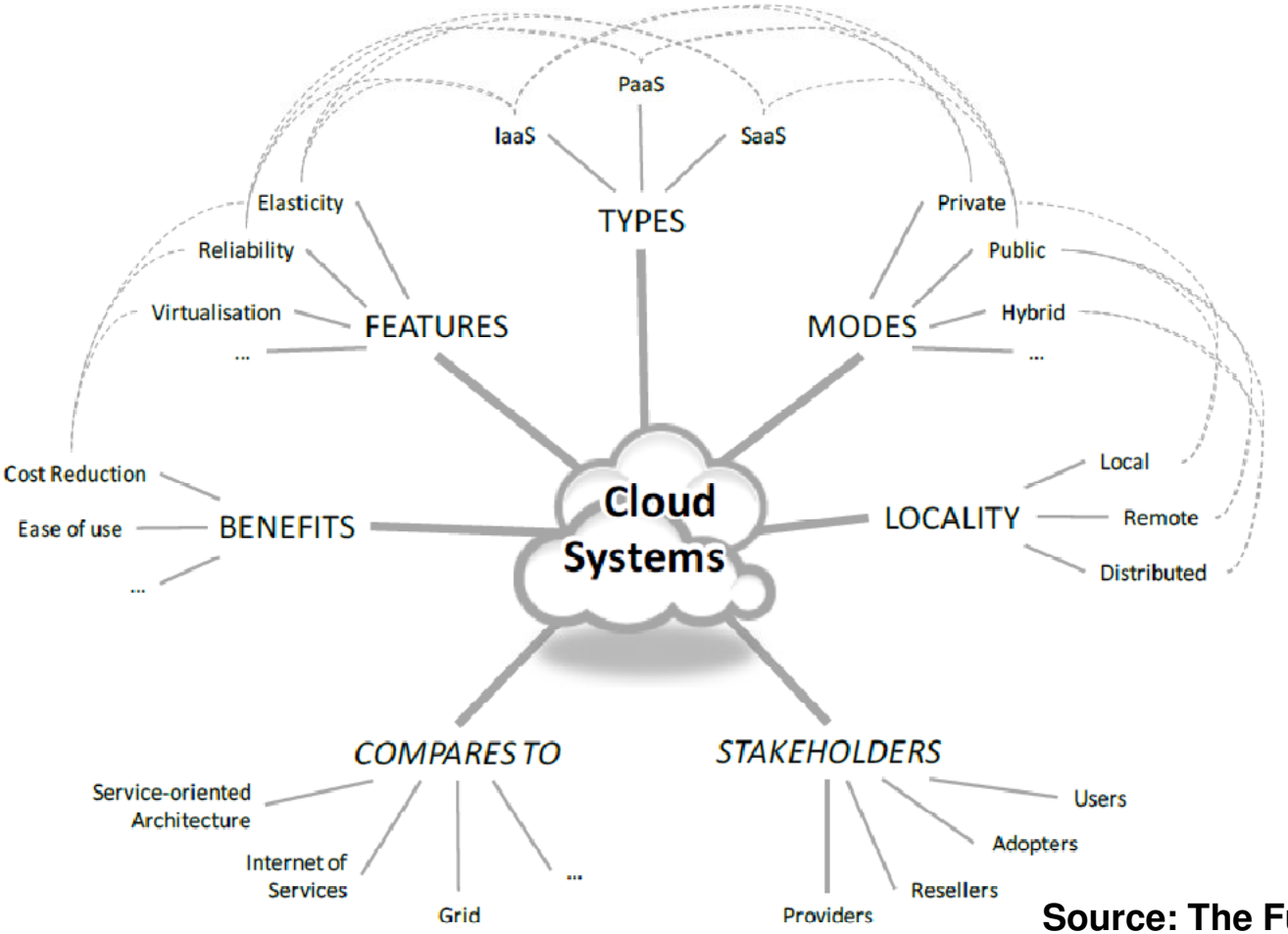
OASIS wg on Identity over the Cloud

SimpleCloudApi, promoted by Zend tech, with IBM,MS et al

Cloud Security Alliance

Open Cloud Manifesto

The Cloud Concepts



Source: The Future of Cloud report

FIGURE 1: NON-EXHAUSTIVE VIEW ON THE MAIN ASPECTS FORMING A CLOUD SYSTEM

From a testbed to a production ecosystem

Oct .'04 | Nov.'07 | Jan.'08 | Oct .'09 | Dec.'09 | Sept.'11

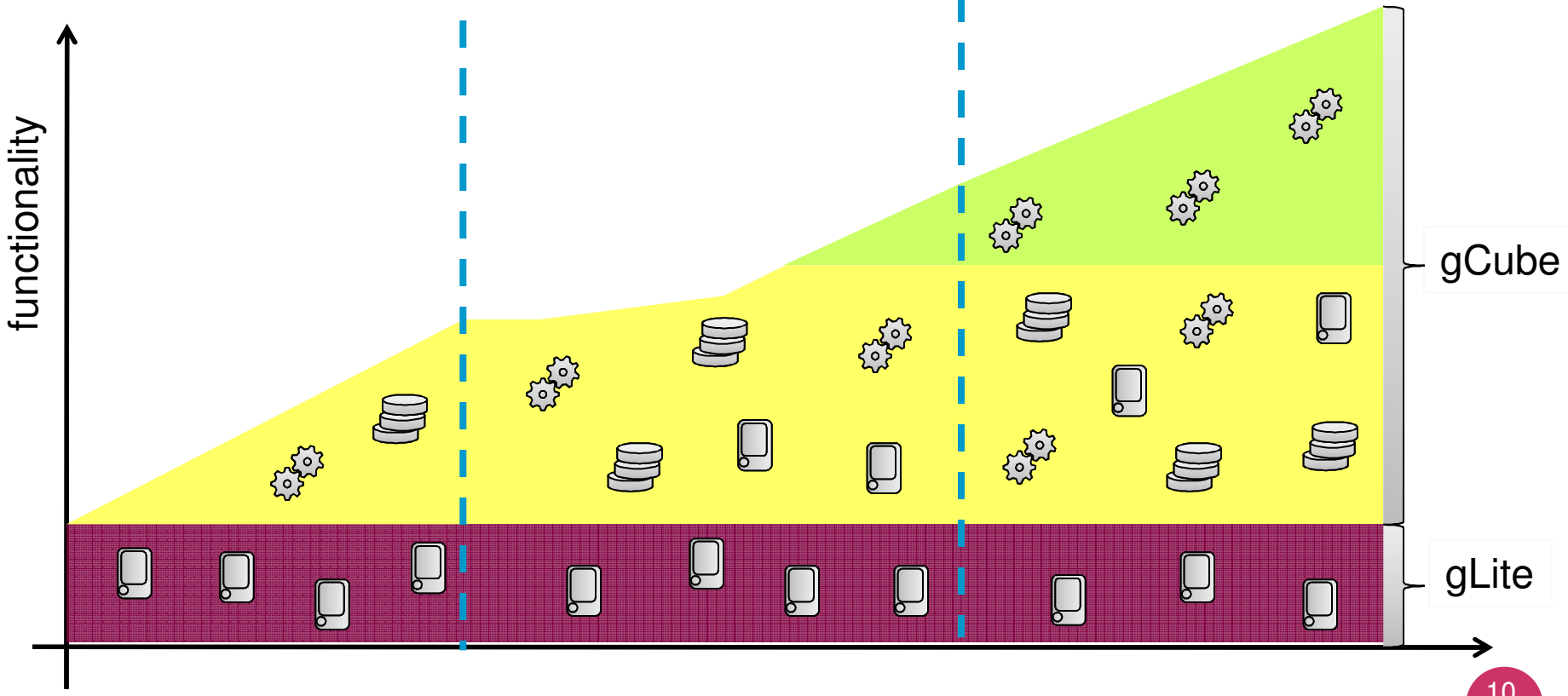
Diligent → D4Science → D4Science II

	Diligent	D4Science	D4Science II
	Testbed	Production	Production
	Empower the grid middleware to: > **manage data** and **metadata** as primary resources > **virtualise** the VO environment	**Stabilize gCube** by supporting two large user communities: > FARM > EM	**Promote interoperability** across e-Infrastructures by empowering large user communities
	Prototype	Software Framework	Open Platform
	=> gCube 0.9	=> gCube 1.6 (stable and open source) => d4science e-Infrastructure	=> gCube 2.0 (feature reach and interop.) => d4science ecosystem

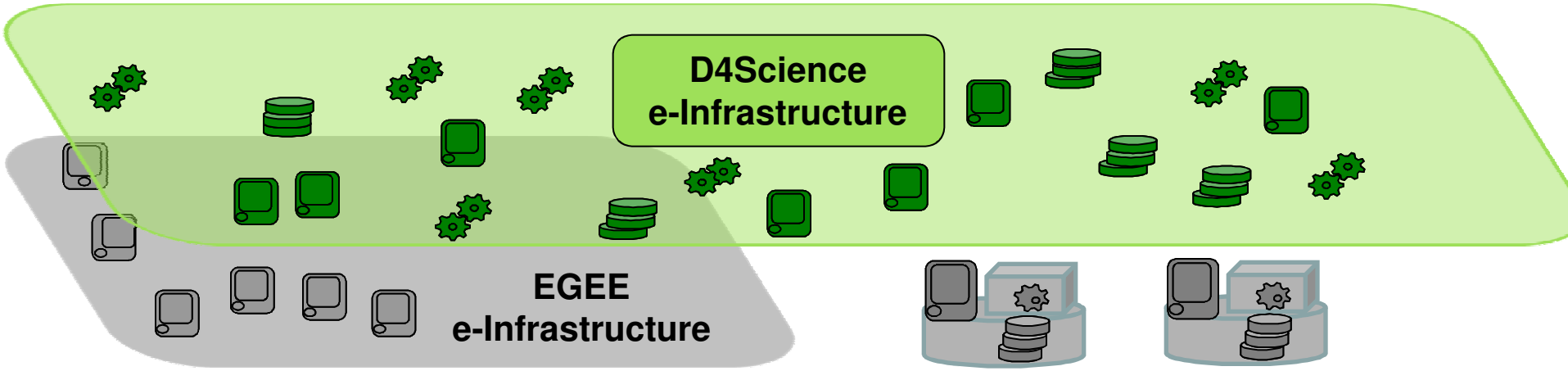
From a testbed to a production ecosystem

Oct .'04 Nov.'07 Jan.'08 Oct .'09 Dec.'09 Sept.'11

Diligent D4Science D4Science II



Sharing



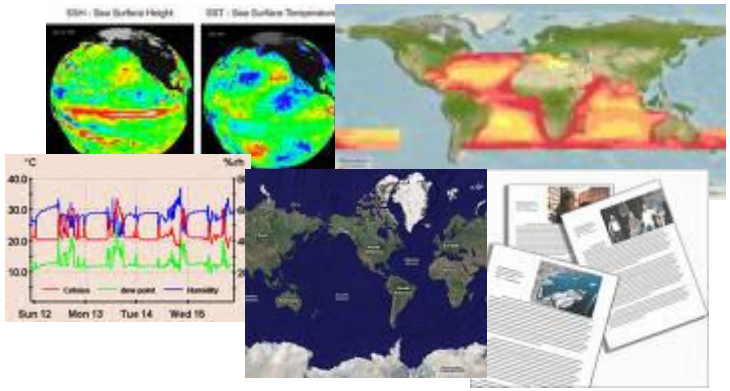
Controlled sharing of

- data resources
- service resources
- computing & storage resourses

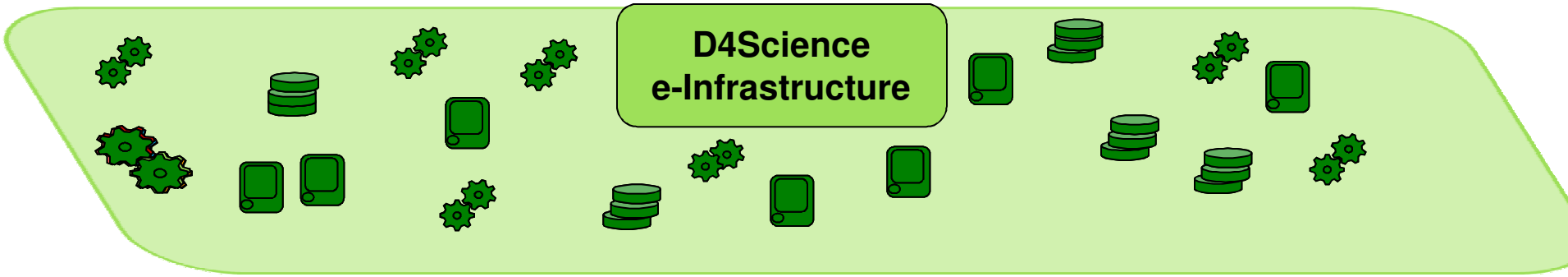
Courtesy – Donatella Castelli

Current D4Science data

- metadata
- environmental reports
- satellite images
- earth observation products
- raw data
- fact sheets
- time series graphs
- country maps
- species distribution maps
- etc.



Data related services



Generic services for:

- search
- annotation
- process management
- visualization
- ……
- time series management
- report generation
- shared workspace
- ……

Courtesy – Donatella Castelli

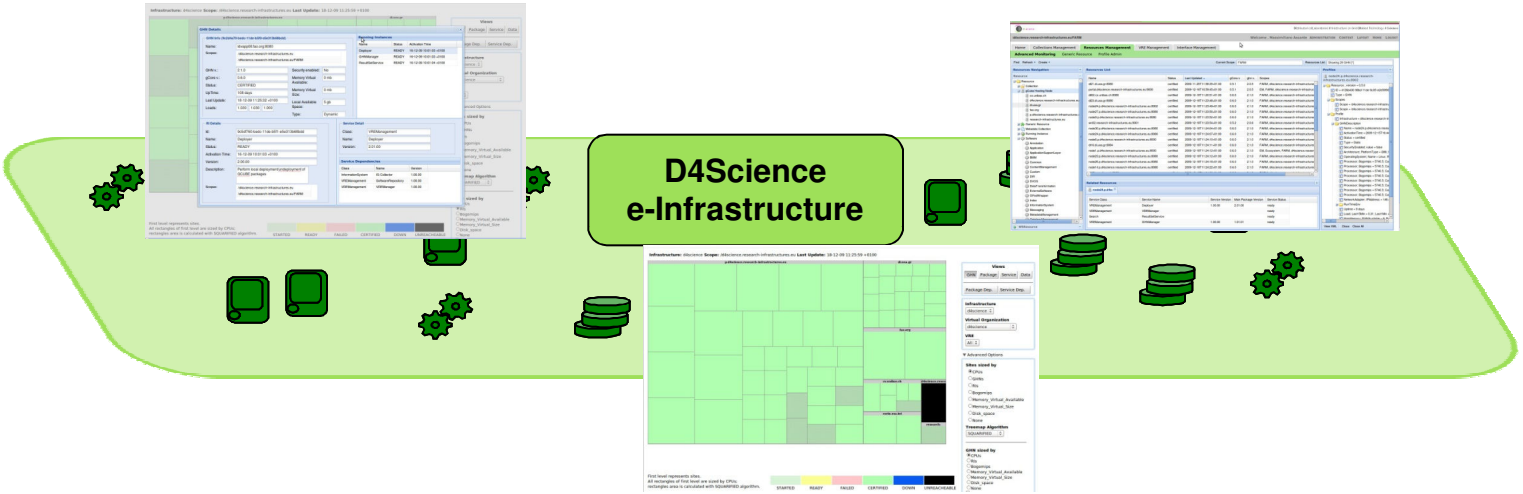
On-demand VREs

Courtesy – Donatella Castelli

Infrastructure administration services

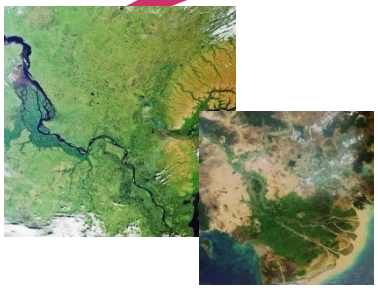
Infrastructure administration services:

- resource registration and management
- user roles management
- infrastructure monitoring
- ……



Courtesy – Donatella Castelli

D4Science User Communities & VREs



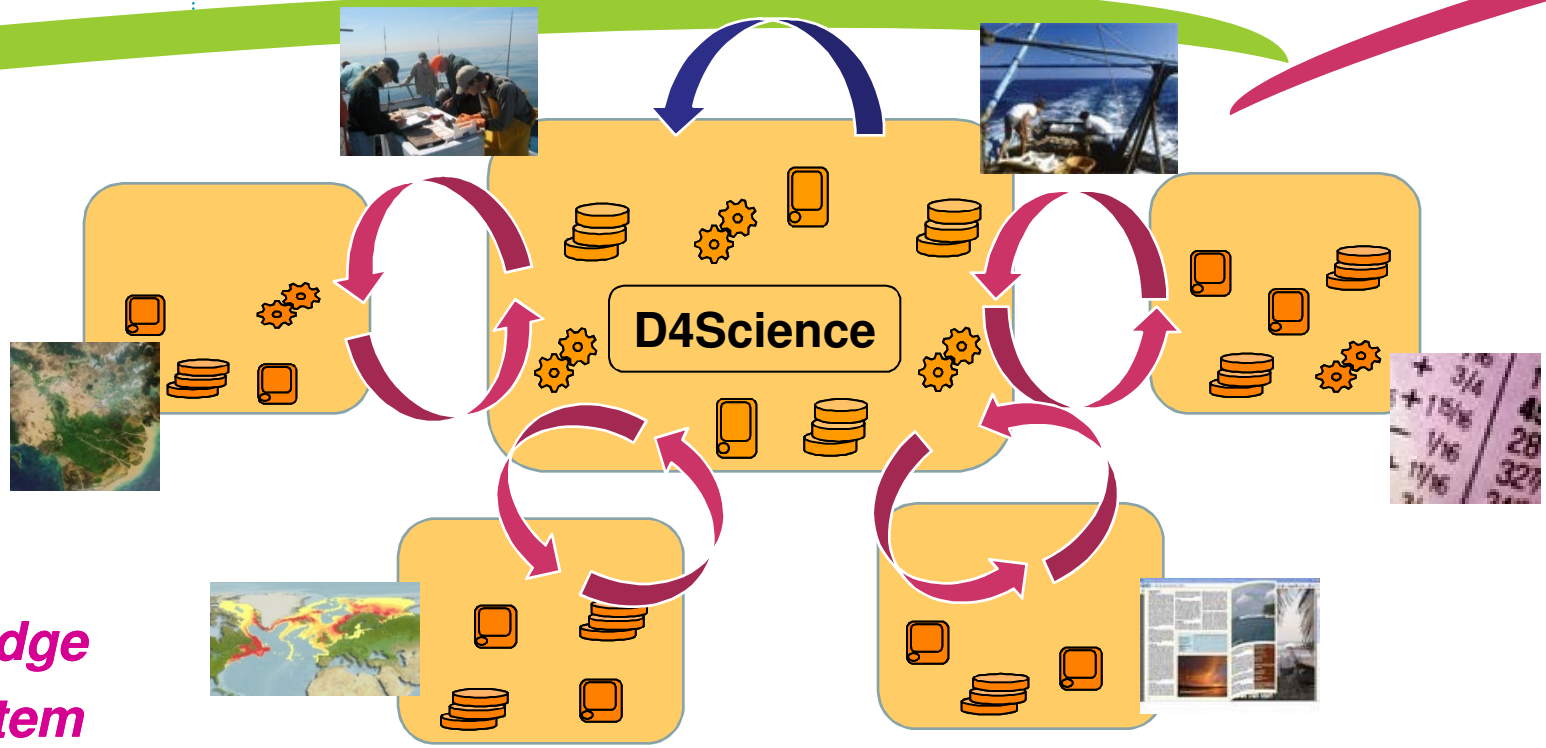
- **Environmental Monitoring Community**
  - Global Chlorophyll Monitoring (GVM) VRE
  - Global Vegetation Monitoring (GCM) VRE

- **Fisheries and Aquaculture Resource Management Community**
  - Fisheries Country Profiles Production System (FCPPS) VRE
  - Integrated Capture Informatio ICIS) VRE
  - AquaMaps VRE

Courtesy – Donatella Castelli

D4Science –II

Knowledge infrastructure ecosystem



Knowledge Ecosystem

- **Interoperability - extend D4Science by introducing mechanisms for facilitating:**
  - the use of data resources managed by different data infrastructures
  - the programmatic exploitation of the D4Science

Courtesy – Donatella Castelli

gCube: enhanced Grid platform

	Grid	gCube
Resources	Computing, Storage	Computing, Storage, Software, Service, Libraries, Plugin, …
Provision	Federation	Federation but dynamic acquisition is supported
Ownership	Decentralized	Decentralized but autonomic
Security	Pervasive	Pervasive but Hidden
Application Model	Task - Oriented	Resource - Oriented
Deployment Model	Static	Dynamic and autonomic
Development Platform	Task-Oriented Libraries	Service Programming Abstraction

Source - P.Pagano, F.Simeoni

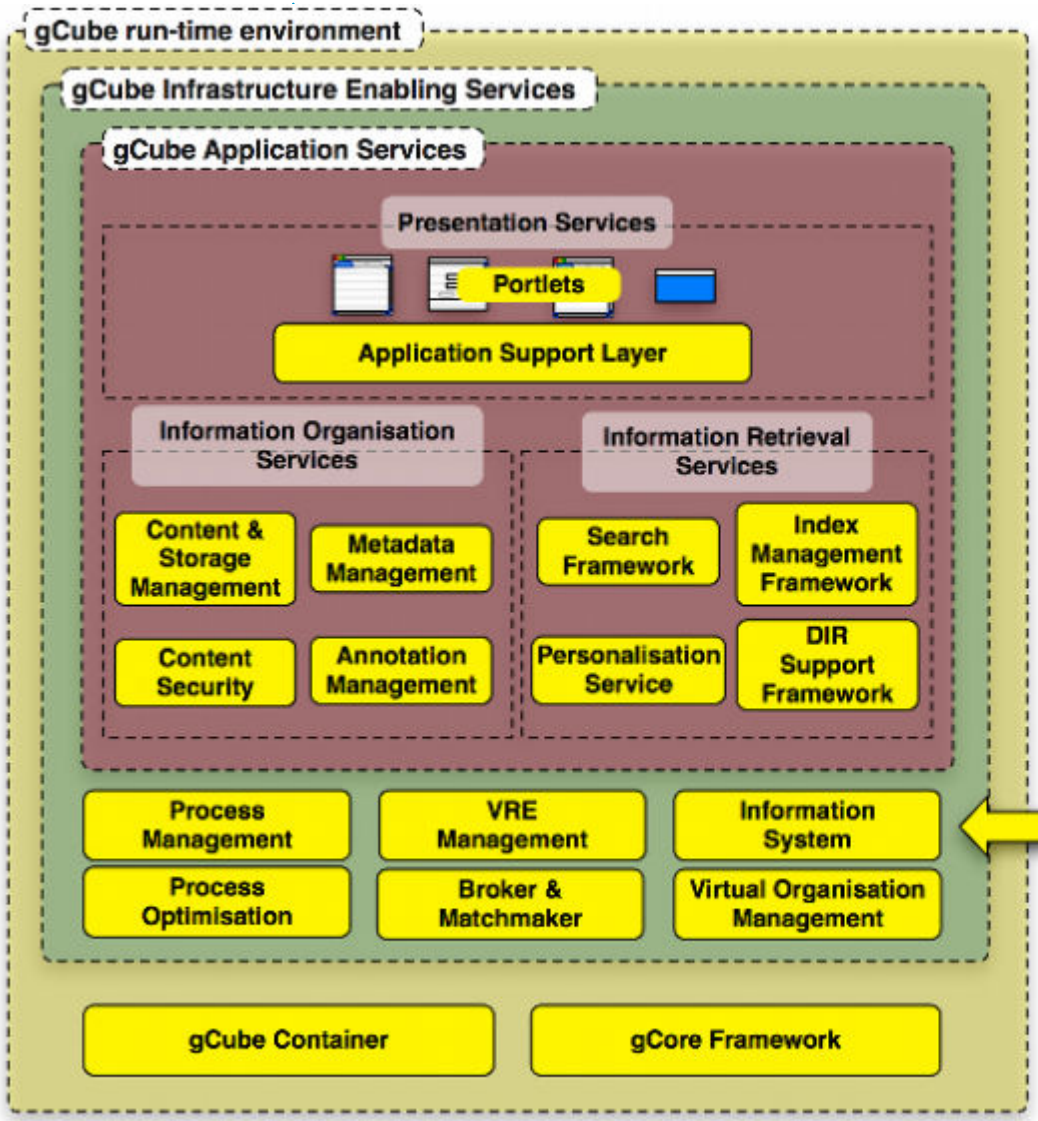
The structure of the study

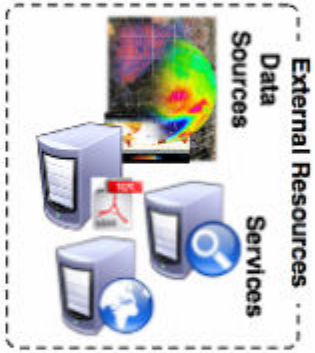
- **Phase 1: Market Analysis and positioning**.
- Expected results:
  - short introduction to cloud market, different options and commercial opportunities
  - gCube positioning with respect the market

- **Phase 2: gCube installation and performance analysis.**
- Expected results:
  - list of possible improvements in the system
  - list of bottlenecks and constraints on the conceptualisation and architecture
  - list of performance indicators and actual capabilities of the sw
  - report on costs for installing and operating a gCube service

- **Phase 3: Analysis of gCube architecture with respect cloud technology**.
- Expected results:
  - report on how gCube may exploit a IaaS
  - report on how gCube may exploit a PaaS.

GCube Architecture





gCube highlights

- Consolidation of physical (computational) resources by adopting virtualisation
  - gCube infrastructure model compatible with IaaS models
- Deployment model similar to Hybrid cloud
  - potential outsourcing to public clouds but still within and towards a federated provision
- gCube as PaaS for vertical market
  - e.g. Salesforce or Google Apps
  - gCube development model analogous to PaaS models compile-time compliance enables runtime management but limitedly to general-purpose services (vs web apps)
- Cloud-level transparencies over grid-style infrastructure

Developing Scientific Data applications

- Exploring alternative paths: VENUS-C
  - A project under negotiation
  - A testbed for Scientific Clouds
  - Focus on federation of clouds and DCIs at platform level

- Exploit Clouds to respond to pick of computational needs.
  - Resources at reasonable costs
  - Integration of on-demand computing resources with the data access and integration capability offered by gCube - increasing the value and usability of such data.
  - Explore issues related to data movement and limitations on Computing far from data repositories.

Sustaining the SDI technologies

- Scientific Data Infrastructures may benefit from clouds
  - outsourcing the storage and preservation of data
  - Investing only on data and paying on the basis of access needs and performance requested
  - Bringing data near the computation

- Some open issues more relaxed but are equally urgent
  - Security, privacy and compliance
  - Vendor lock-in
  - Data cost curve may bring a reduce in overall investments

Toward a Scientific Cloud for Data

- Parade (the Partnership for Accessing data in Europe)
- Caspar (FP6 Project) Usage of Cloud in long term preservation
- Vision on technologies for Cloud Storage
- www.dataliberation.org promoting open data format for clouds

- The right time for Cloud4Science - Storing, Preserving, Accessing and Discovering data for e-Science

Conclusions

- Cloud – a successful marketing concept that boosted an innovative use of existing technologies
- D4Science Ecosystem, is providing PaaS-like capabilities to build Virtual Research Environments and will explore
  - how to evolve gCube to meet technological challenges imposed by Cloud offers
  - costs and benefits of porting a Scientific Data Application on top of (federated Cloud Platform provided by VENUS-C)

- Why not a “Cloud4Science initiative”?

This is not the end…

- What will happen when an entry-level server (now bi/quad processors dual cores) will be 64 cores (or more) on a single processor?
- What will happen to your (our) solutions?
- Will Virtualisation technology still the solution?
- Will we need for new programming languages? And translators for existing languages?

A long term roadmap for e-Infrastructures and related technologies (middlewares and services) should start to take into account multicore evolution.