



d4SCIENCE

**Deploying e-Infrastructures to enable
multidisciplinary collaboration in
Environmental Science**

Barcelona, 22 September 2009

**Virtualizing research environments to hamper the complexity:
D4Science solution for data sharing, manipulation, and analysis**

Pasquale Pagano (CNR – ISTI)
D4Science Technical Director
pasquale.pagano@isti.cnr.it



- D4Science narration
- Empowering the e-Infrastructure
- Hampering the complexity
- Considerations

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

Diligent

D4Science

D4Science II

Testbed

Empower the grid middleware to:

- > manage Data and metadata as primary resources
- > virtualise the VO environment

=> gCube 0.9

Production

Stabilize gCube by supporting two large user communities:

- > FARM
- > EM

=> gCube 1.2 (stable and open source)

=> d4science e-Infrastructure

Production

Promote interoperability across e-Infrastructures by empowering large user communities

=> gCube 2.0

=> d4science ecosystem

From a testbed to a production ecosystem

Oct. '04

Nov. '07

Jan. '08

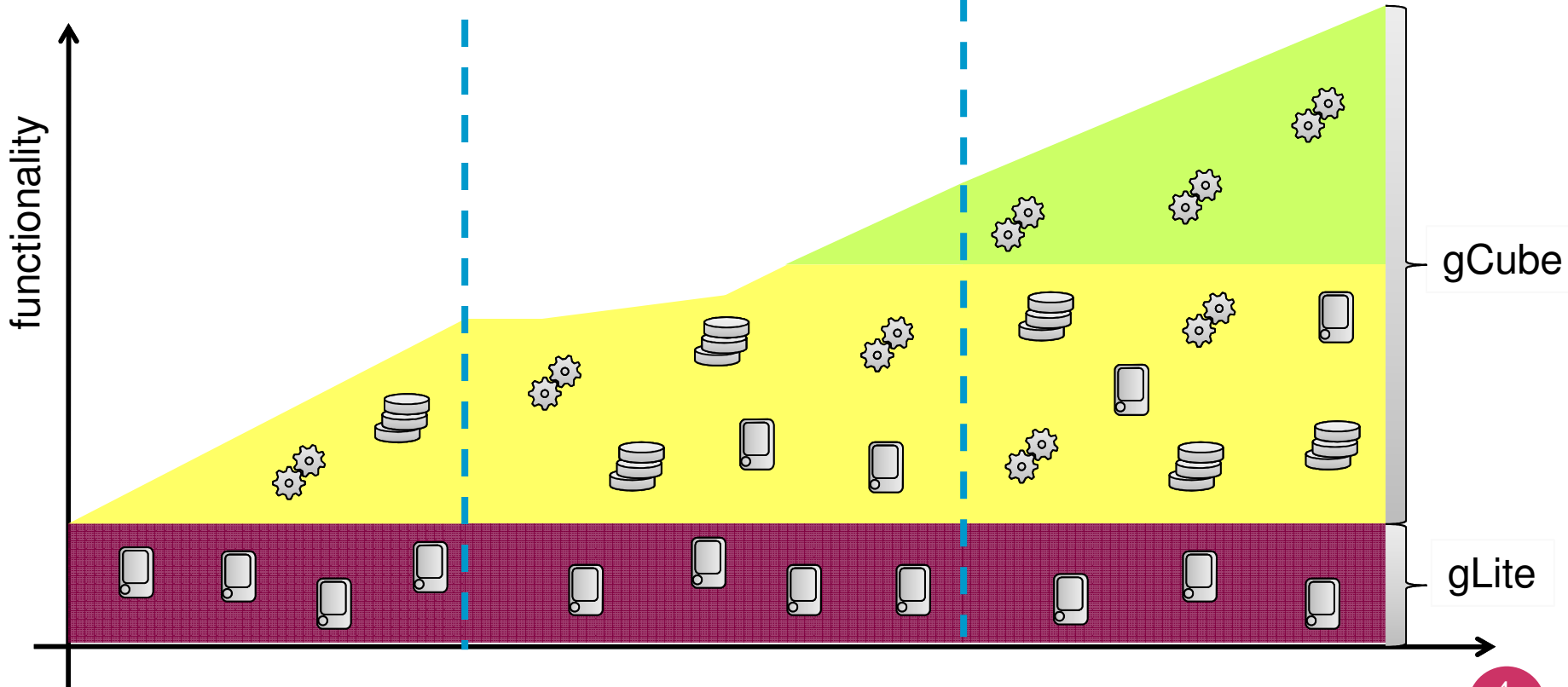
Oct. '09 Dec. '09

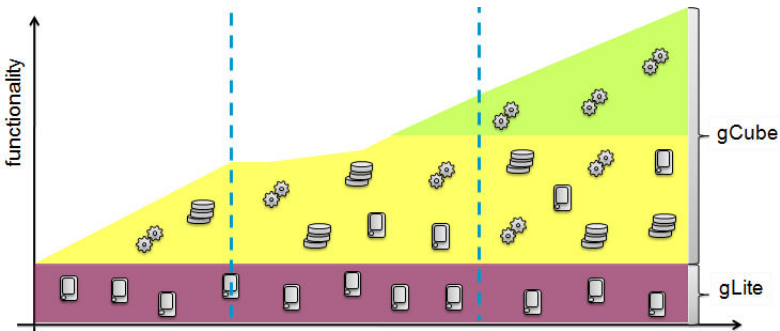
Sept. '11

Diligent

D4Science

D4Science II



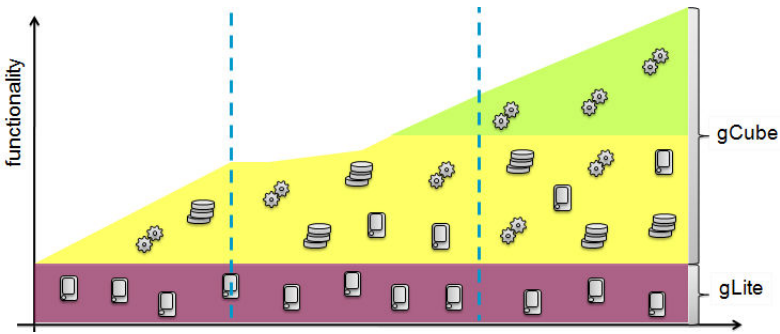


Is a simple Registry sufficient to manage such growing set of heterogeneous resources?

➤ NO, it is not !

gCube provides an **Information and Monitoring System** where rich set of resources including computing, storage, service, data, metadata, and applications can be:

- registered, discovered, and accessed
- monitored, shared in a controlled way, accounted



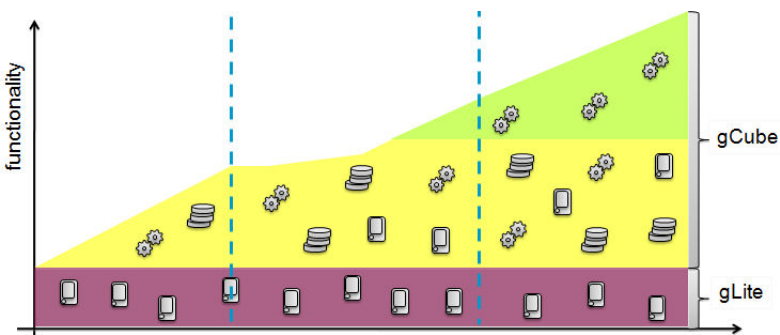
Is the VO adequate to represent such growing aggregation of resources tailored to satisfy needs of the members of the community?

➤ NO, it is not !

Data needs to be assessed before to make it publically exploitable by the VO members.

Restricted set of users have to collaborate to refine processes and implement show cases.

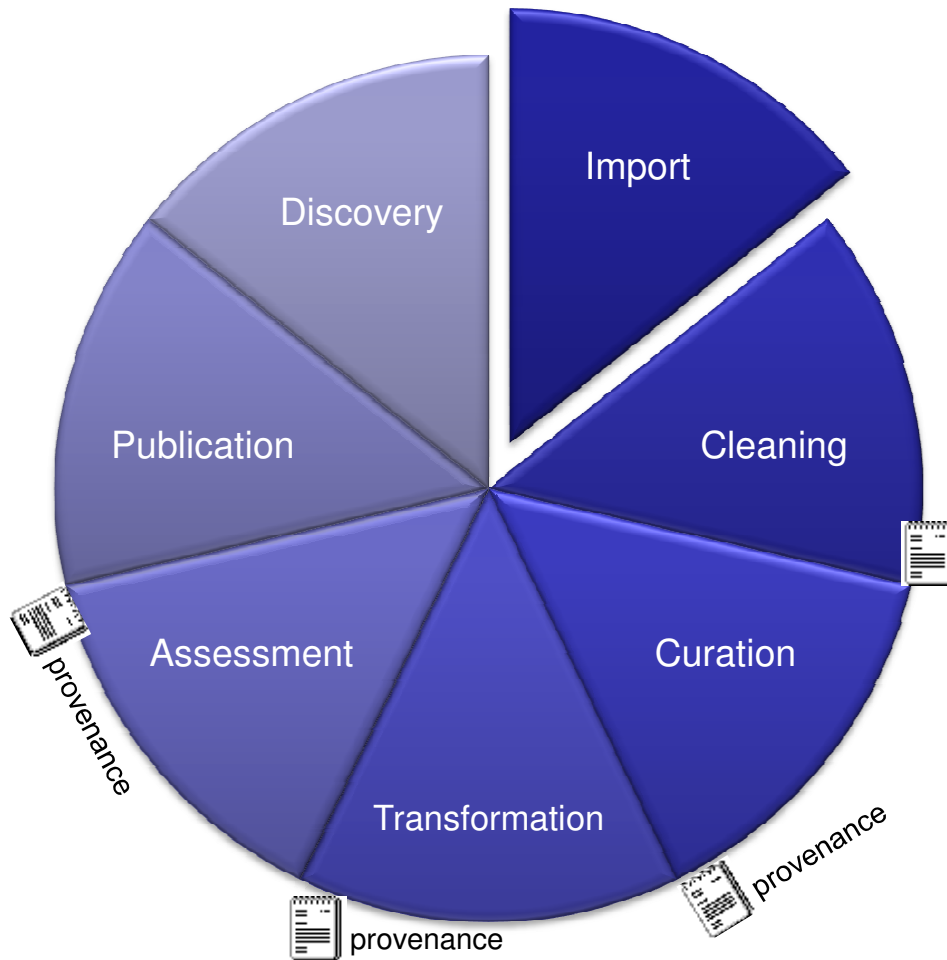
Products generated through elaboration of data or simulation have to be validated by expert users.



gCube provides an environment where subset of resources can be assigned to a subset of users for a limited timeframe through a **Virtual Research Environment (VRE)**

VRE allows to collaborate with other members of the VRE through a shared VRE workspace where data, processes, and products can be commented and annotated.

VRE resources can be published in the VO at any time by the VRE data managers.



- Data and metadata resources lifecycle is extremely complex
- gCube VREs provide support for all steps through **plug-in based services**
- All steps are reproducible since **provenance metadata** are automatically generated



Data Sharing

Import + Cleaning

Acquisition/linking data from SE, desktop, database, Web, GridFtp, ..

Generation/linking metadata,

Validation of the input, definition of data types

Publication

Sharing with a user-defined set of collaborators

Sharing with all members of the VRE

Sharing with all members of the VO

Discovery

By full-text and geo-spatial search; by exploring any collection of data

By taxonomies and ontology

By subscription and notification

Data Manipulation and Analysis

Curation

Tailored set of applications for XML metadata (any schema) and relational data

Support the enrichment of data and metadata through the exploitation of reference data

Transformation

Transformation of data and metadata performed by:

- Using existing VO applications
- Registering tailored VRE applications
- (transparently) clustering data and executing jobs on the infrastructure

Assessment

Performed by set of experts through:

- Comparison of generated and existing products (for metadata and relational data)
- Collaboration with colleagues
- Annotation and metadata enrichment

Live resource repository

- Common representation of heterogeneous resources
- Common model to support publication, monitoring, and accounting
- Runtime enforcement of resources policy

Virtual desktop

- Specialised research environment
- Created on demand by assigning resources to users
- Fine-grained resource sharing

Controlled data sharing

- Fully-integrated support for data and metadata lifecycle
- Support for heterogeneous metadata and data representations

Controlled end-user applications support

- Secure sharing of end-users transformations and applications
- Facilitate the integration, usage, and exploitation of tailored applications