Data infrastructures Interoperability and Standards: the DRIVER and D4science experiences

OGF Digital Repositories Workshop

Wednesday 4 March 2009, 16.00-19.00, Catania, Italy, within OGF25

Pasquale Pagano
CNR-ISTI
Pasquale.pagano@isti.cnr.it

Driver and D4Science

- Two FP7 projects
- Commonalities
  - Deliver a Service Oriented Infrastructure
  - Adopt common design principles and design patterns
  - Adopt many common standards
- But
  - Rely on two different software framework (D-Net and gCube)
  - Are partially interoperable

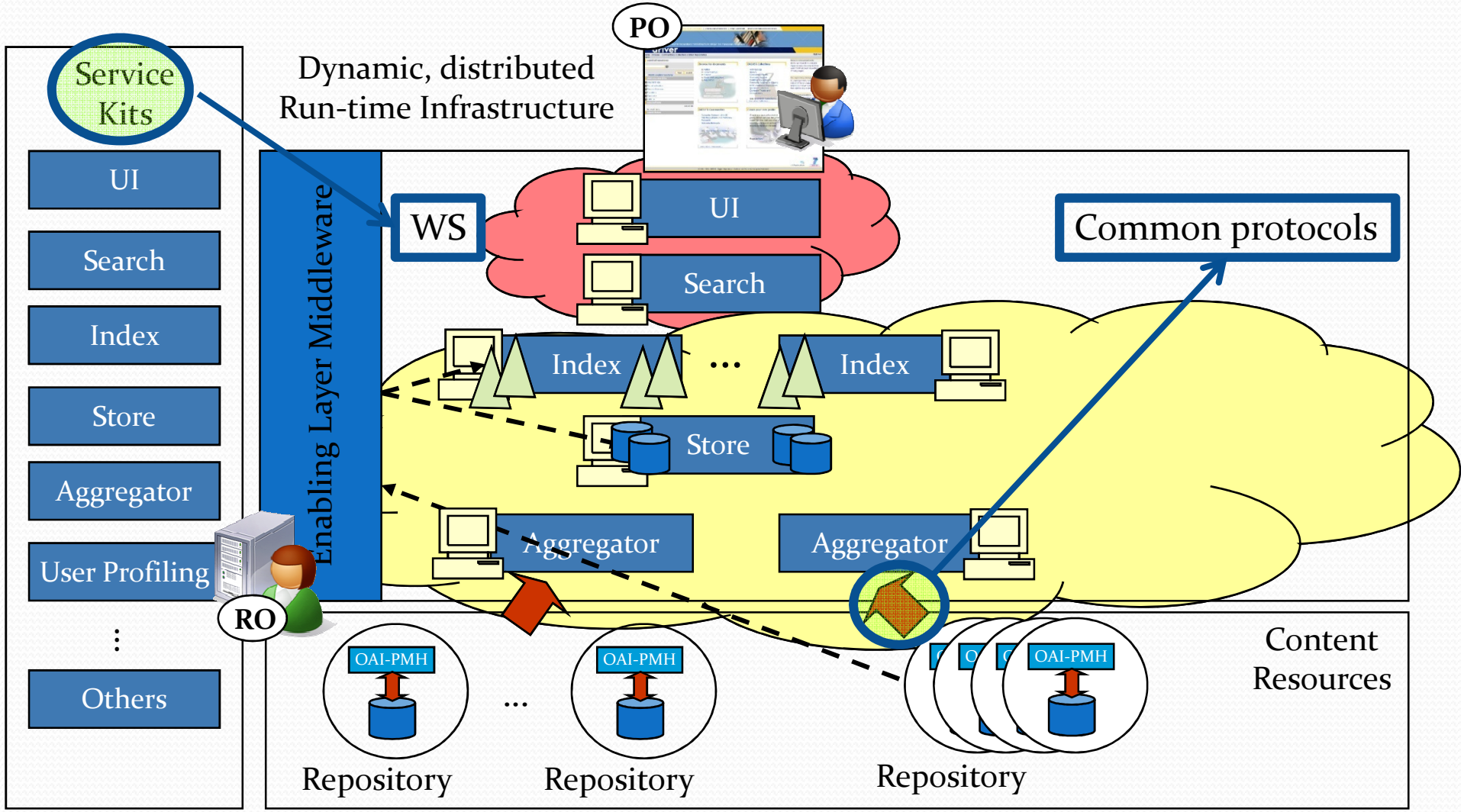
WHY?

Driver and D4Science [cont.]

- Driver was born to
  - Realize an **European Information Space** for Open Access research publications
  - Provide demanding research communities with free and easy access to this Information Space through User portals and through applications and services
- D4Science was born to
  - Enable the creation of **Virtual Research Environment:** framework of applications, services and data sources dynamically identified to support the underlying processes of research, collaboration, and cooperation
  - Remove technical concerns from the minds of scientists, *hide all related complexities from their perception*, and enable them to focus on their science and collaborate on common research challenges while exploiting Grid infrastructure

Is this enough to explain the differences?

Driver in a nutshell



Courtesy by Paolo Manghi

Towards standard: Driver guidelines

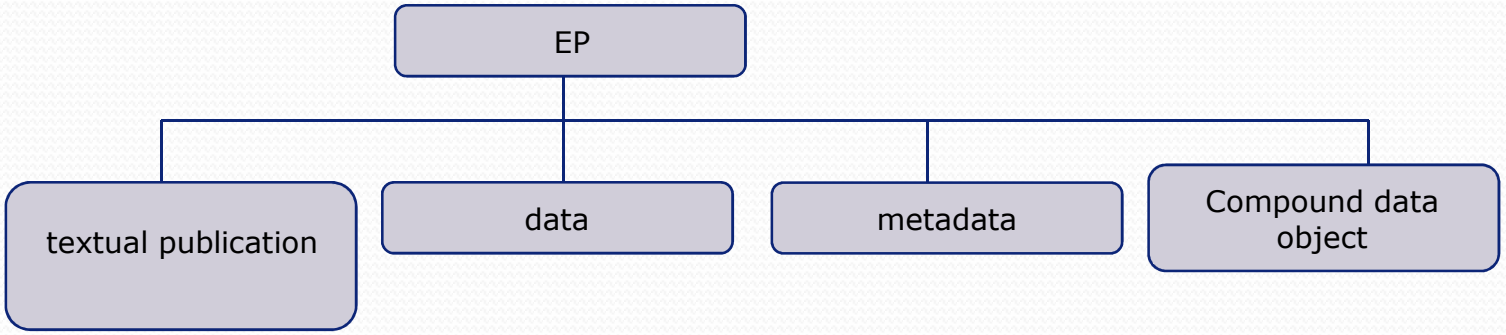
- Guidelines for Repository Managers and administrators on how to expose digital scientific resources using OAI-PMH and Dublin Core Metadata, creating interoperability by homogenizing the repository output.

Does your repository follow the DRIVER guidelines?	n	%
We do not know about the DRIVER guidelines	49	27.5
We know about the DRIVER guidelines, but do not follow them	32	18.0
We know about the DRIVER guidelines and (make every effort) to follow them	97	54.5

Courtesy by Driver Consortium

Towards standards: Driver Enhanced Publication

- An Enhanced Publication (EP) is:
  - a textual publication enhanced with:
    - research data (evidence of the research) and/or
    - extra materials (to illustrate or to clarify) and/or
    - post-publication data (commentaries, ranking)
  - So: ever evolving



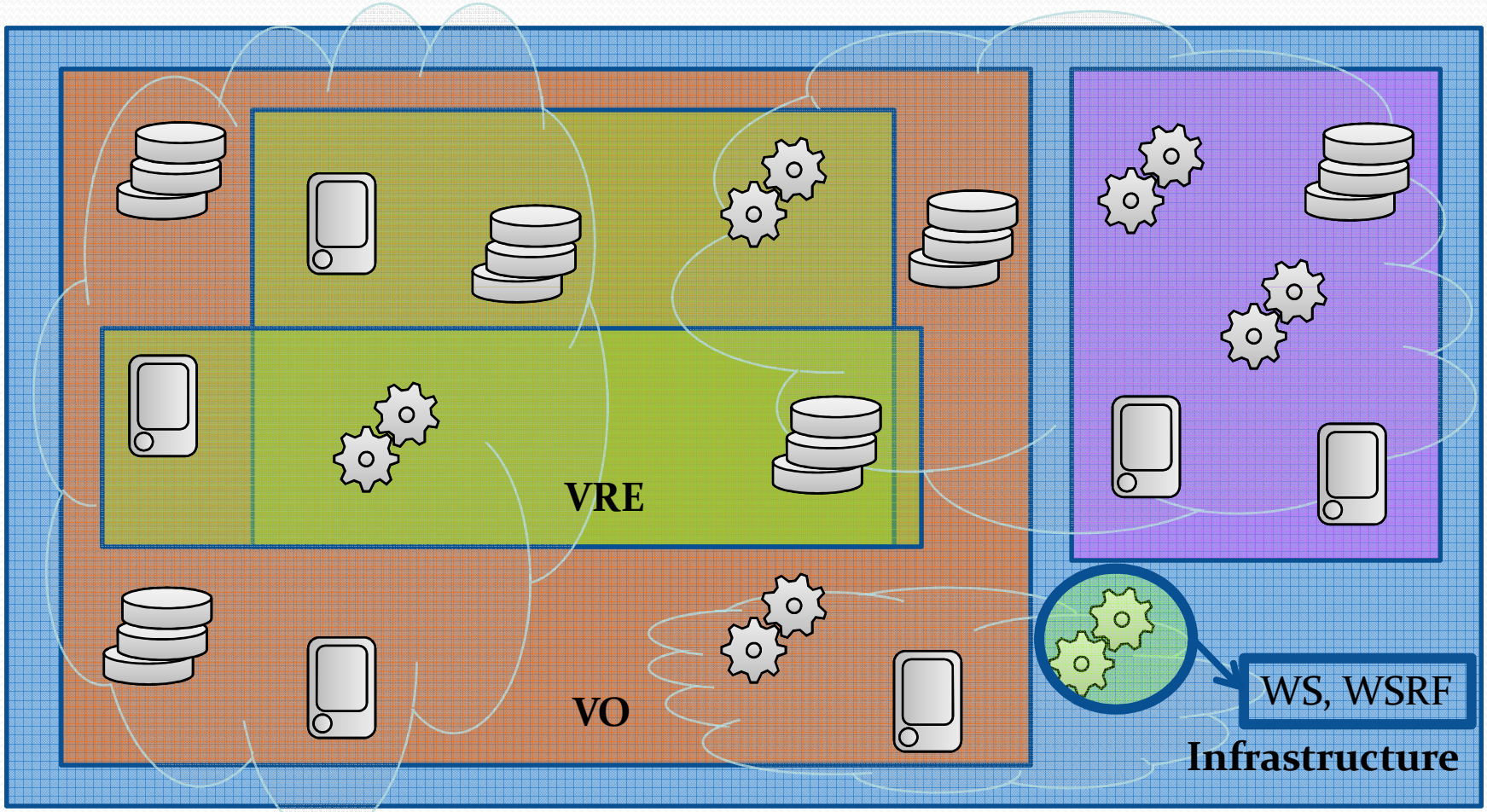
Courtesy by Driver Consortium

Driver production infrastructure

- Content
  - More than 186 repositories (more than twice to come) over 21 countries have been integrated
  - More than 856,264 open access documents
- Services
  - Currently 25 different kinds of Services
  - Production release: 36 service running instances over 9 nodes located at CNR, ICM and NKUA
- Applications
  - DRIVER Main, Belgium, Spain-Recolecta

Courtesy by Paolo Manghi

D4Science in a nutshell



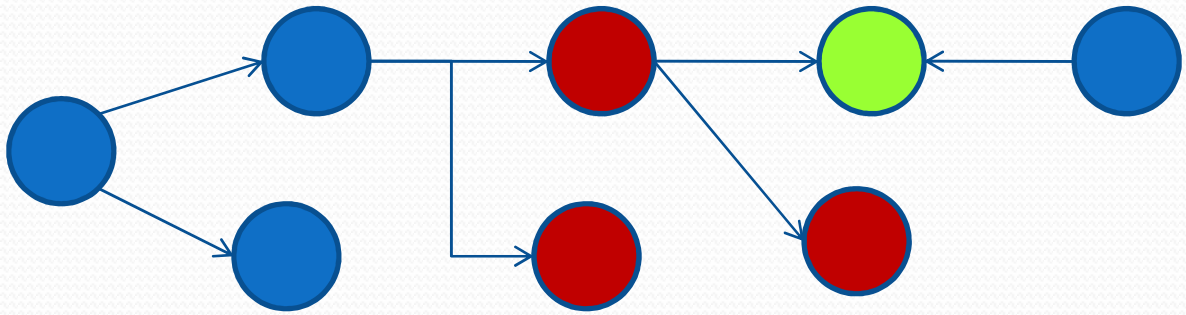
Towards standards: gCube resource profiles

- Many types of resources: computing, storage, data and service
  - Glue Schema for describing Computer, Storage, and Grid components
  - gCube Service Profile for describing WS interfaces, build and runtime dependencies, deployment constraints, deployment preferences, software packages
  - gCube Data Profile for describing provenance and tailored VO information

Towards standards: gCube compound object model

- SRM provides a complete interface to heterogeneous storage systems
  - consistent homogeneous interface to the Grid, while allowing sites to have diverse infrastructures.
  - Born taking into consideration actual use cases and influenced by needs of the large High Energy Physics communities
  - Capabilities: directory and ACL, non-interference with local policies, space reservation and management, abort, suspend, and resume operations, transfer protocol negotiation, …

- gCube embeds SRM (and GridFTP) support by providing an higher interface allowing to store, discovery, and access qualified network of files (**compound objects**).



D4Science infrastructure

- Content
  - 37 data sets (more than twice to come) maintained by international organizations, e.g. ESA, FAO, WorldFish Center
- Services
  - Currently 59 different kinds of Service and 290 software packages
  - More than 500 WS running instances distributes on 4 sites consuming 245 GB RAM - 26,666 TB - 87 CPUs
- Applications
  - **Fishery Country Profiles Production System (FCPPS)**
  - **Integrated Capture Information System (ICIS)**
  - **Global Ocean Chlorophyll Monitoring (GCM)**
  - **Global Land Vegetation Monitoring (GVM)**
  - **…**

Conclusion

- Standards are not enough to secure infrastructure interoperability
- But they are needed

- Resources representation: repeating the Glue Schema experience? Extending the Glue Schema?
- Compound objects management: defining SRM +?
- Guidelines and best practices for the adoption of the specification