Ανοιχτή Πρόσβαση – Γνώση για Όλους
Αθήνα, 16/12/2008

Ηλεκτρονικές Υποδομές Ψηφιακών Αποθετηρίων Περιεχομένου & Υπηρεσιών

Γιάννης Ιωαννίδης
Πανεπιστήμιο Αθηνών

Outline

- Concepts
- The Driver Projects
- The DILIGENT / D4Science Projects

Concepts

- Digital Repository
  - ✓Storage and basic retrieval services
- Digital Library
  - ✓Advanced retrieval services
  - ✓Mostly documents
- Virtual Research Environment
  - ✓Value added services, e.g., collaboration & computation services
  - ✓Data and documents
- Scientific Data Infrastructure
  - ✓Multiple VREs
  - ✓Cross-domain collaboration

DRIVER I & II

DILIGENT

D4Science

…

eInfrastructures

- Collection of electronic resources for domain-specific services in multiple application domains
- Critical for environments requiring
  - ✓Vast distributed pool of physical & virtual resources
  - ✓Distributed & cross-organization access to / ownership of resources
- Virtual Organizations are key for partitioning & sharing resources in an eInfrastructure
  - ✓Can span multiple physical organizations

eInfrastructure Resources

- Hardware and system software (operating sys …)
- Enabling services (connectivity, power)

GEANT

- Enabling software (middleware …)

EGEE

- Content (datasets, documents, ontologies …) *DRIVER*
- Generic and application-specific software

DILIGENT, D4Science

- Humans (users, administrators …)
- Policies (auth/access, protocols …)

DRIVER

- Digital
- Repository
- Infrastructure
- Vision for
- European
- Research

Scholarly Communication: Imperatives

- Comprehensive, global access to any type of scientific information
- Minimum time and resources effort to access and use this information
- Easy search/navigation, handling, manipulation, and re-dissemination of information
- Maximum visibility to and communication with the research community, research impact
- Long-term access and preservation of research results

(European) Open Access Vision

Berlin declaration (2003)

Free and unrestricted access to sciences and human knowledge representation worldwide

All research institutions in Europe and worldwide make all their research publications openly accessible through institutional and thematic repositories.

DRIVER Vision and Objectives

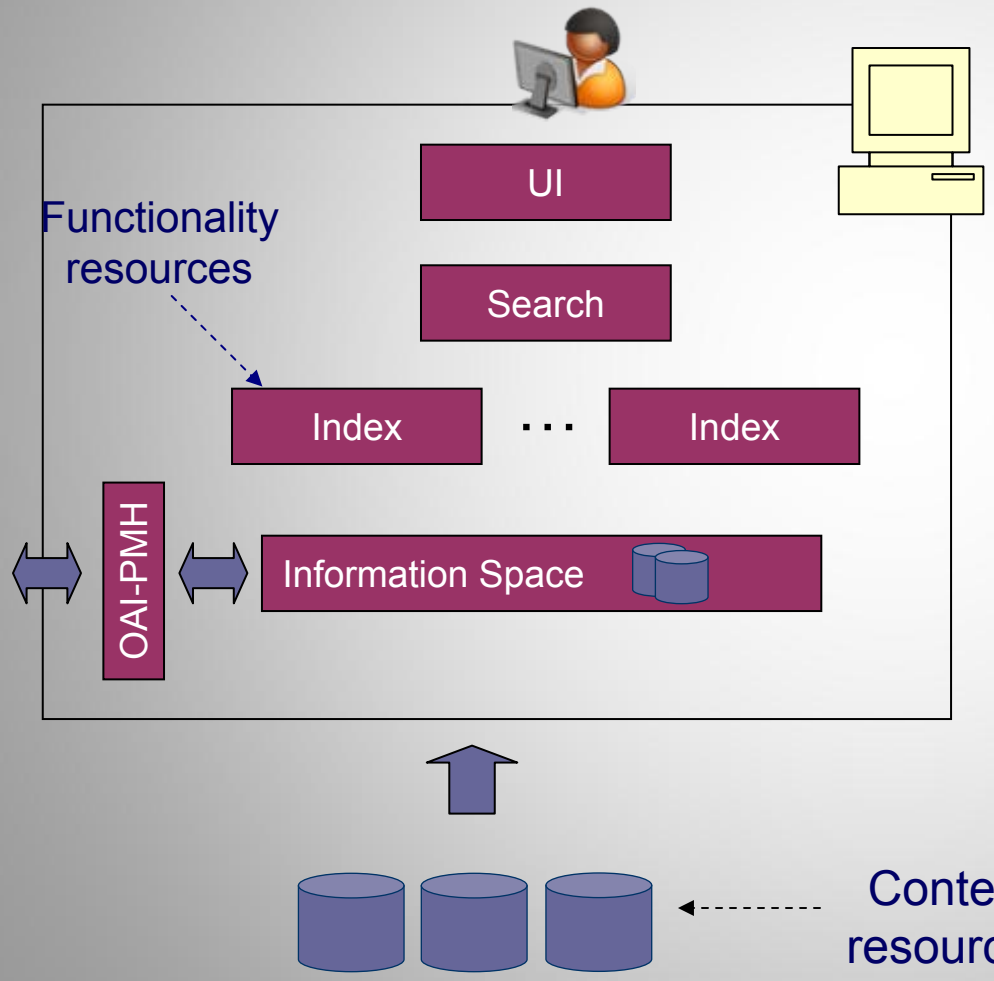
- Environment for integrating existing national, regional, or thematic repositories
- Production-quality European DR infrastructure
- Future expansion and upgrade to ***the*** European DR infrastructure
- Identification & promotion of relevant standards
- Raising awareness among user communities

Driver Partners

- University of Athens (GR) - coordinator
- University of Bielefeld (DE)
- CNR-ISTI (IT) – technical management
- SURF Foundation (NL)
- Univ. of Nottingham – SHERPA (UK)
- University of Bath – UKOLN (UK)
- University of Warszawski – ICM (PO)
- University of Gent (BE)
- University of Goettingen (GE) – scientific management
- Danish Technical University (DN)
- National and University Library (SL)
- University of Minho (PT)

Repository Systems: current efforts

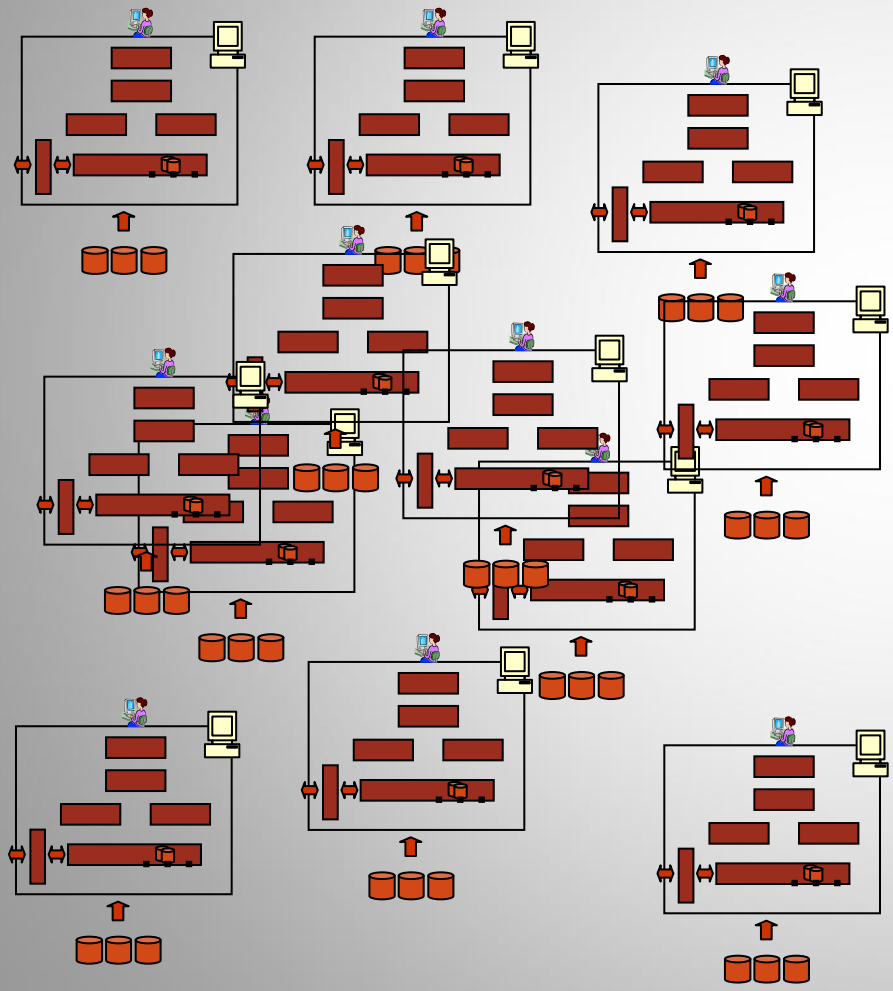
Individual institution site



- Centralized System
- High hardware & software installation & maintenance cost
- Poor & limited scalability
- Reuse by data and service duplication!

Repository Systems : current efforts

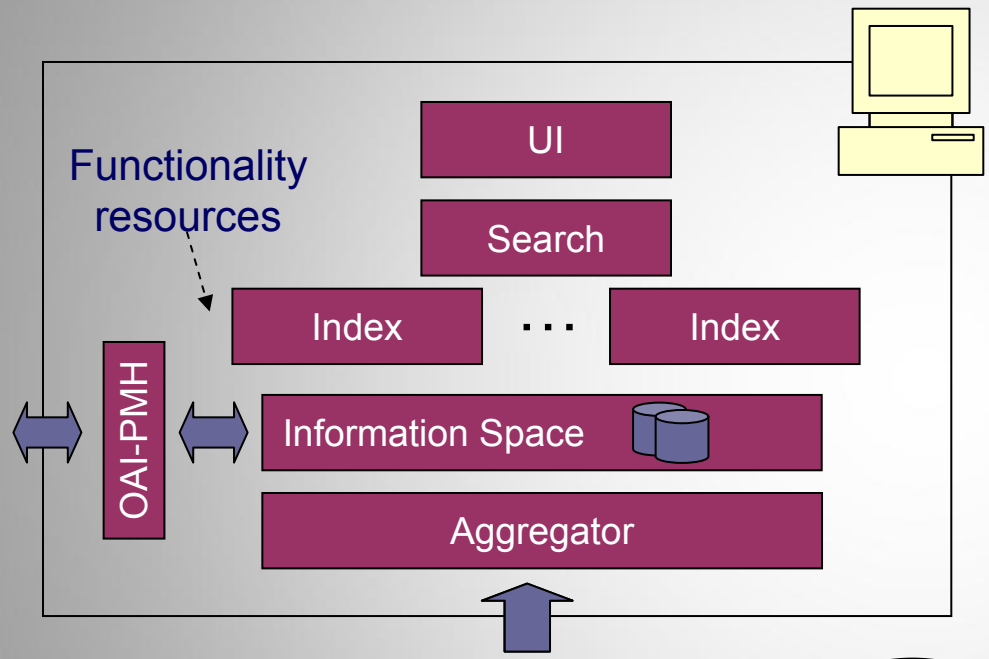
Multiple institution sites

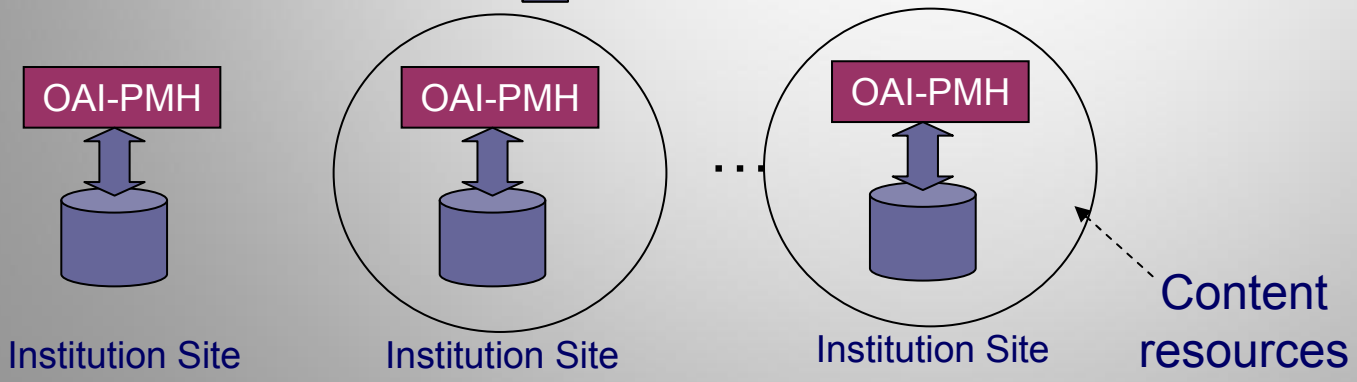


- Repeated efforts
  - ✓High hardware & software installation & maintenance cost
  - ✓Poor & limited scalability
  - ✓Reuse by data and service duplication!
- Disconnected repositories

Repository Systems : current efforts

Sharing and reusing content

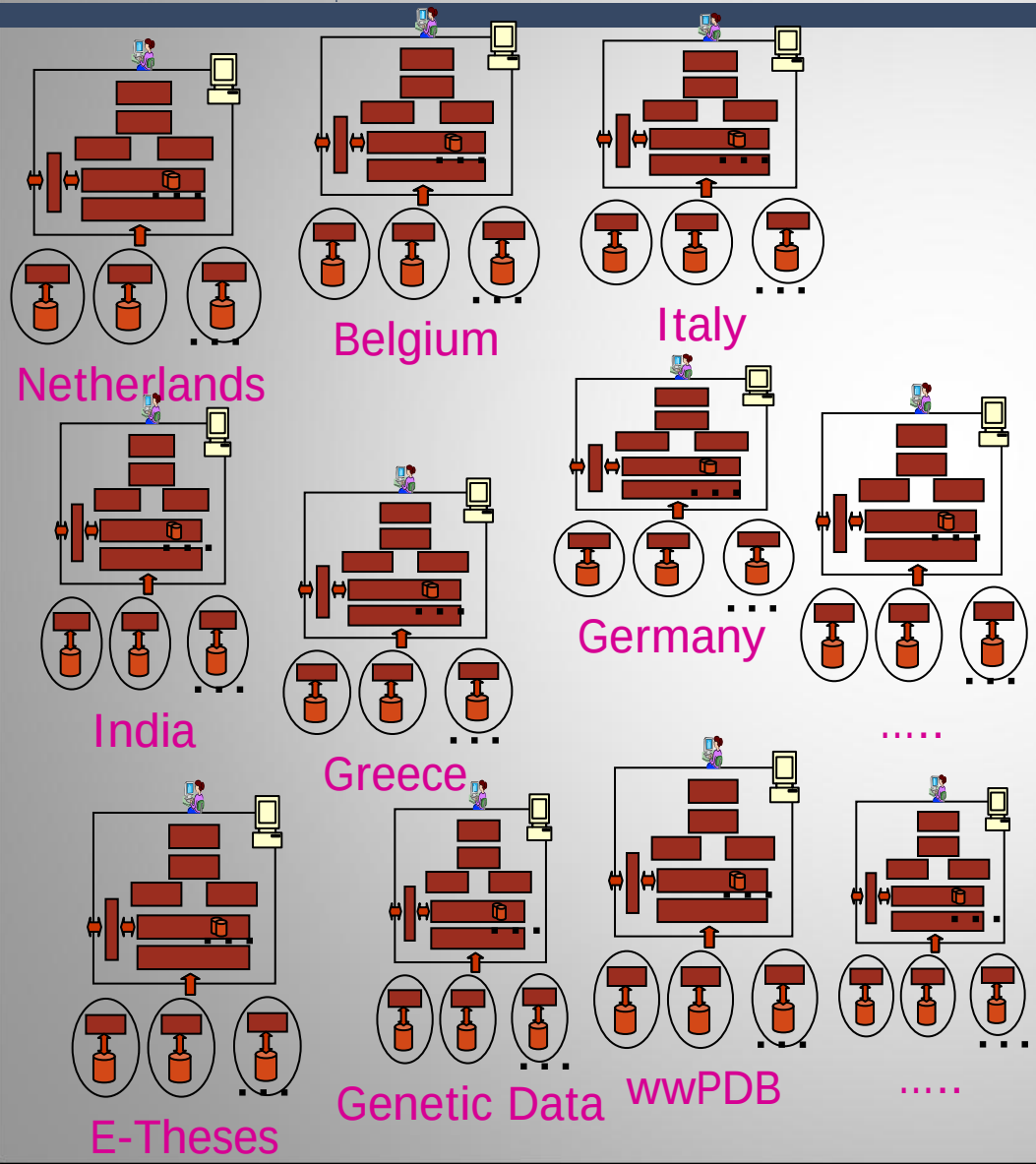




- Centralized System
- High installation and maintenance cost for hardware and software
- Poor & limited scalability
- Reuse by data duplication!

Repository Systems : current efforts

Sharing and reusing content



- Repeated efforts
  - ✓High installation and maintenance cost for hardware and software
  - ✓Poor & limited scalability
  - ✓Reuse by data and service duplication!
- Disconnected repositories
  - ✓Sometimes desired policy
  - ✓Often undesirable

DRIVER Infrastructure Vision

Moving from

building individual repositories
or repository clusters,

one at a time,
repeating “things” again and again,

to building

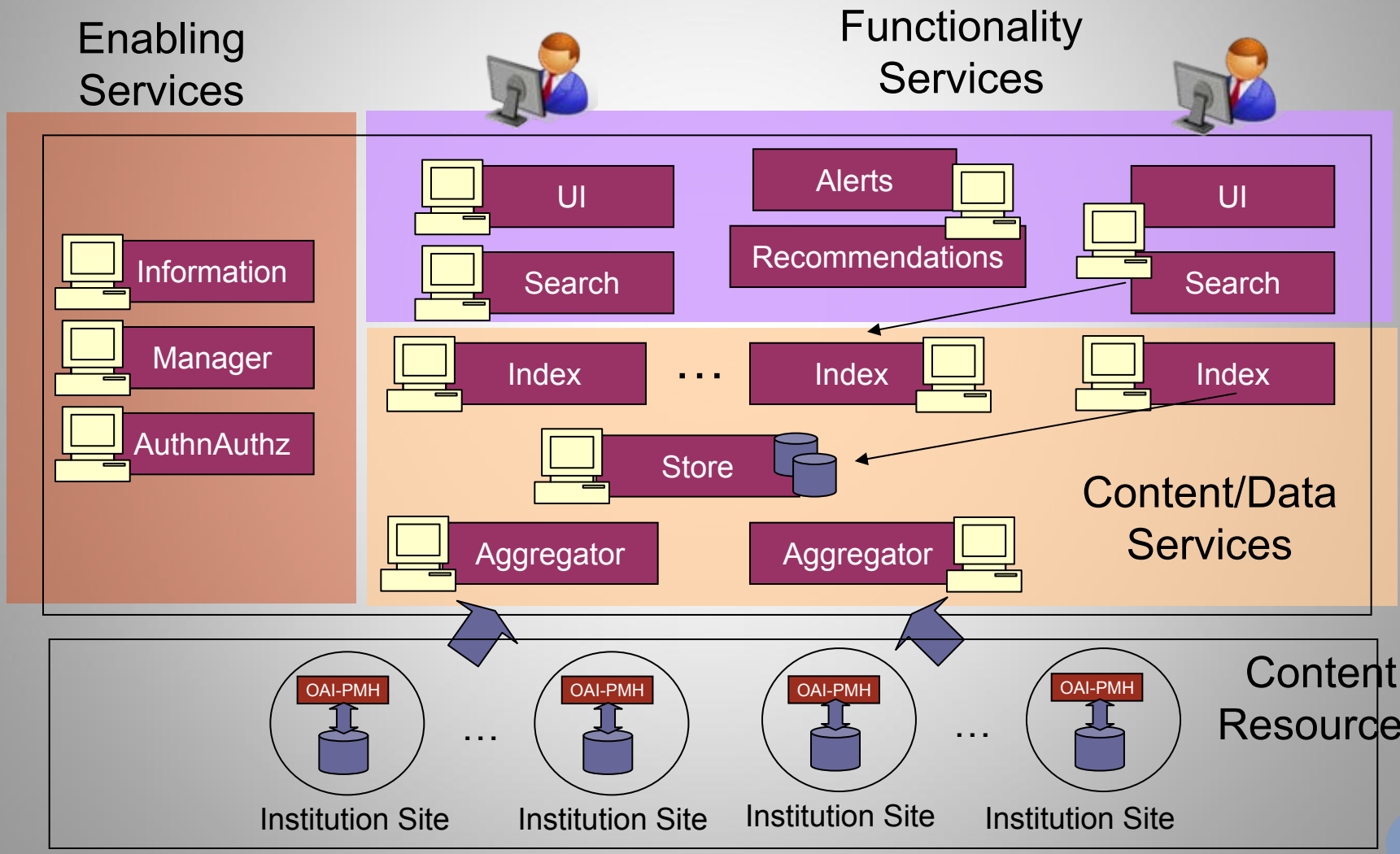
a “generating engine”,

a warehouse,

an INFRASTRUSCTURE,

facilitating the above by offering
appropriate generic, reusable services

DRIVER Infrastructure



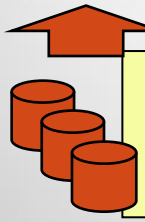
DRIVER Infrastructure: another view

System Managers e.g., DRIVER Consortium

User communities

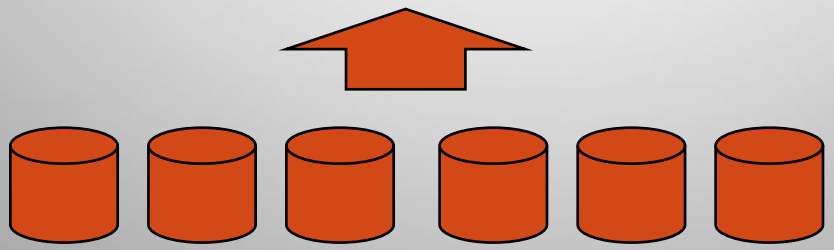
Application ... Application

Service providers, e.g.
DRIVER Consortium
National communities
Subject communities etc.



Aggregation/ Harmonisation system

National communities etc.



Data providers (EU Repositories)

Technological Advantages

- Scalable and dynamic
  - ✓Repositories are dynamically added
  - ✓Scales up with usage/load
- Extensible
  - ✓New functionalities & services are easily added
- Fully Distributed System
  - ✓*Web Services* and Service Oriented Architecture
- D-NET v 1.0
  - ✓First public release of the DRIVER software toolkit

Applications and Uses

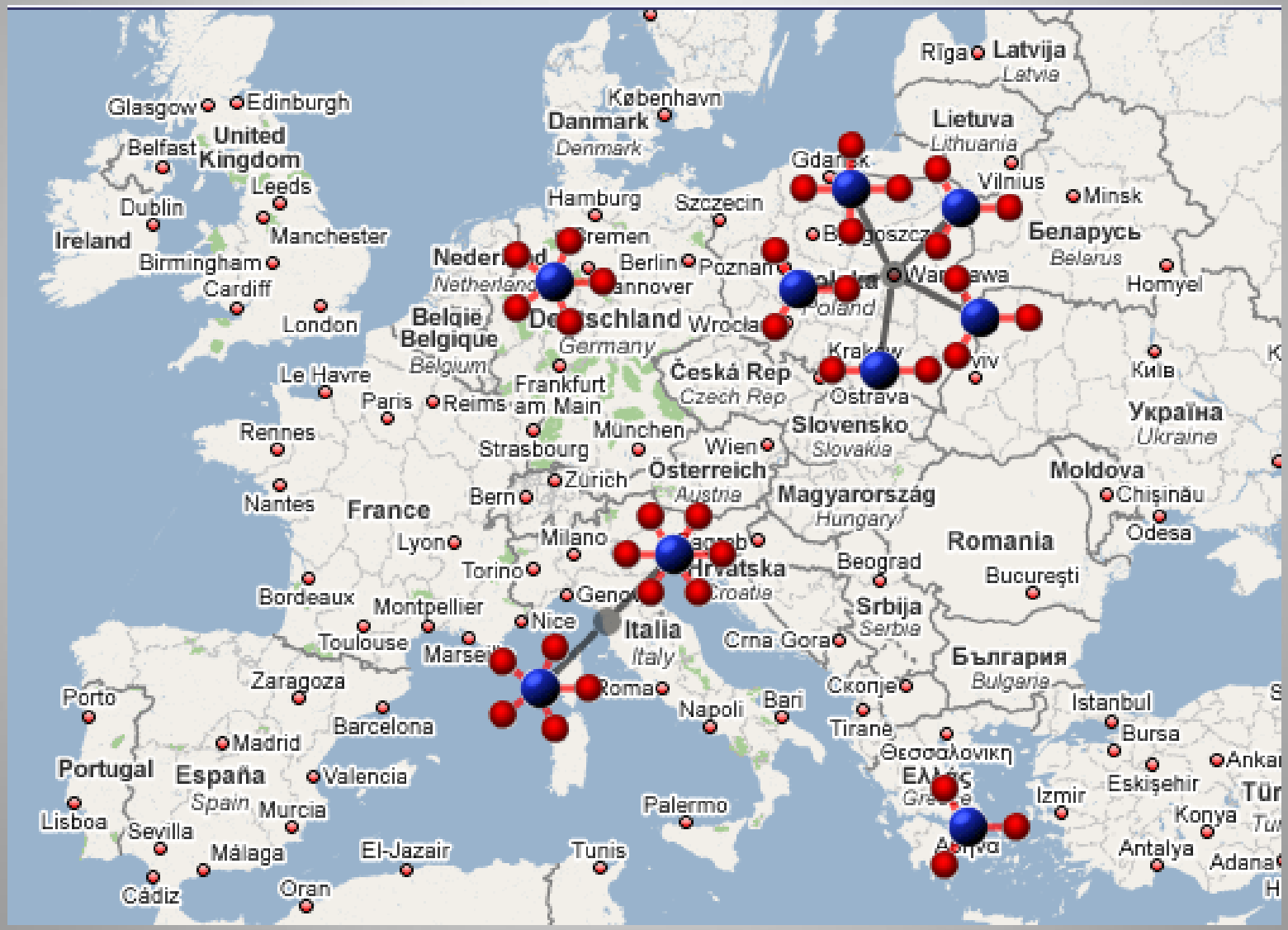
- For researchers
  - ✓Advanced searching capabilities
  - ✓Collections offering specialized views on the content
  - ✓Communities allowing for collaboration
  - ✓User personalization mechanisms
  - ✓Alerts and recommendations
- For repository managers
  - ✓Repository registration and content validation tools
  - ✓Increased visibility through DRIVER portal

Service Providers

Use cases by national, countries, communities, ...

- Joining existing DRIVER instance, e.g., w/ own portal (RECOLECTA, Spain)
- Running own independent DRIVER instance (Belgium)
- Validating own repositories w/ DRIVER Validator

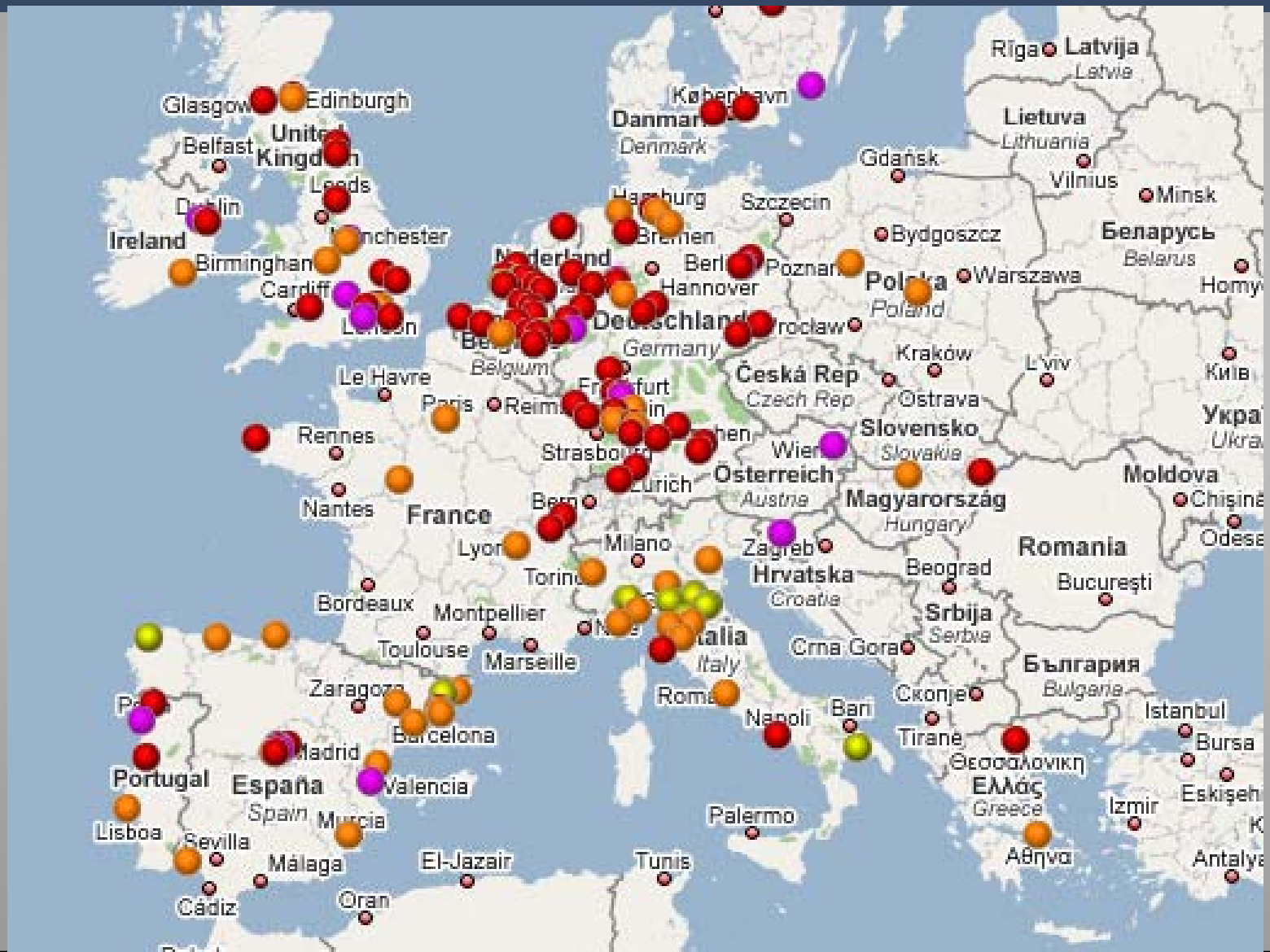
Current DRIVER Instance: hardware/software resources



Current DRIVER Instance: content resources

- 800,000 Open Access documents
- 160+ European repositories
- 15 European countries
- 25+ languages
- 15 document types (research papers, thesis, books, conf lectures, etc.)

Repository Map



DRIVER Connection to National Repository Communities

- National communities represented by country "correspondents"
  - ✓One institution (group), e.g., DARENet-NL, SHERPA-UK, OA-Netzwerk-GE, RECOLECTA-ES, HAL-FR
- Country "correspondents"
  - ✓Maintain national repository information on DRIVER Wiki
  - ✓Organise repository events in own countries
  - ✓Translate repository guidelines and other relevant information into national languages
  - ✓Build up national data aggregators, clean data, offer additional services

DRIVER Connection to Int'l Repository Communities

- Catalyst for global repository infrastructure
- European repository infrastructure node
- Liaison with institutions and initiatives from majority of European countries, the U.S., Canada, Latin America, China, Japan, India and Africa
- MoUs with SPARC Europe, LIBER, eIFL, Recolecta ES, OA-Netzwerk GE, and DRF Japan

DRIVER Confederation of Repository Communities

- Members and strategic partners invited
  - ✓European and international repository communities
  - ✓Subject based communities
  - ✓Repository system providers
  - ✓Service providers
  - ✓Political, research, funding etc. organisations

Future *(D-NET version 2.0)*

- Support different media types of content
- Full text search capabilities
- Provide support for rich publications
  - ✓ Enhanced publications (ORE)
    - Aggregation & discovery of primary data
  - ✓ Processing of data (link to D4Science?)
- Complete advanced functionalities
  - ✓ Communities
  - ✓ Personalized services

More information about DRIVER

- Go to the DRIVER main website
  www.driver-community.eu

- Contact the DRIVER Helpdesk
  helpdesk@driver-support.eu

DILIGENT / D4Science

- DIgital
- Libraries over
- Grid
- Enabled
- Networks
- Testbed

- 4 Science

Virtual Research Environment

A System, comprising of heterogeneous physical and human resources, policies, specifications, software and data / information / knowledge, that enables cooperation and knowledge production in a scientific domain, by offering distributed, cross-organization, facilities for diverse, domain-specific, analysis and processing

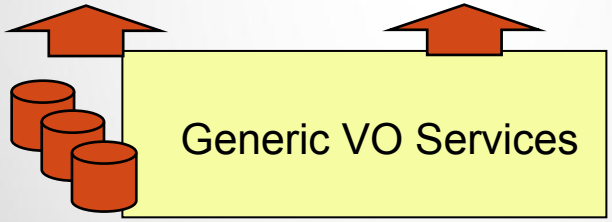
D4Science Infrastructure: on-demand VRE services

System Managers e.g., D4Science Consortium

User communities

Application ... Application

VRE specific Functionality

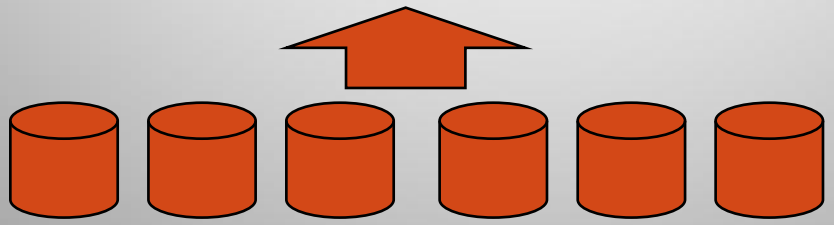


gCube by D4Science (Open Source)

Generic VO Services

GRID middleware gLite

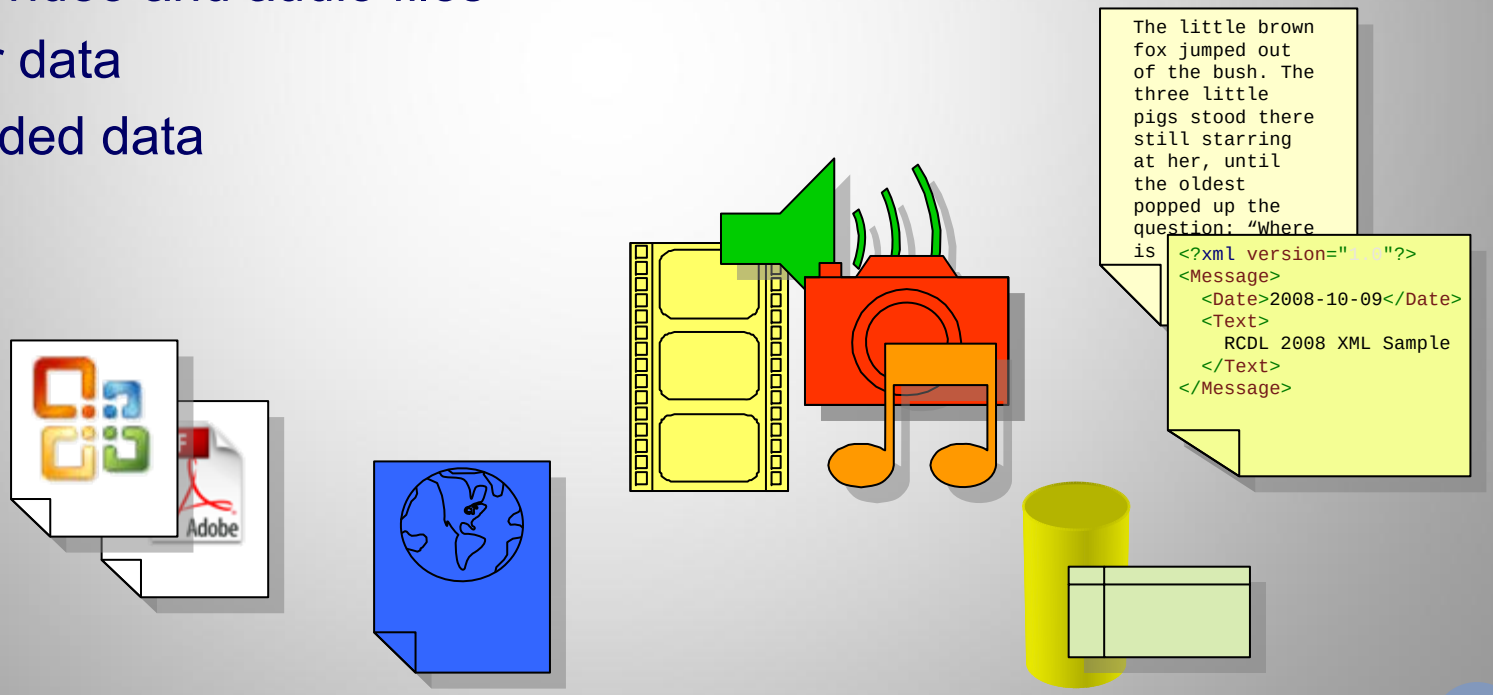
Sci community



Data providers (Sci Communities)

Nature of Data in a gCube eInfrastructure

- Data in gCube: anything that can be stored digitally
  - ✓Plain text files (unstructured, (semi) structured)
  - ✓Binary-form textual document files (pdf, doc, …)
  - ✓Image, video and audio files
  - ✓Tabular data
  - ✓Geo-coded data
  - ✓…



Optimal Use of VRE Resources in Information Retrieval

- Essential for:
  - ✓ Maintaining QoS contracts
  - ✓ Confronting infrastructure-raised challenges
  - ✓ Attracting resources to the Grid
- Special challenges:
  - ✓ Uncontrolled environment
    - Access to resources
    - Access to resource meta-information
    - Abrupt parts & joins
  - ✓ High-dimensional search space
  - ✓ Multi-facet quality metrics
  - ✓ Heterogeneity
    - Resources
    - Meta-information

In gCube

- **Pre-query Optimization**:
  - ✓ Keeper service monitors and adapts the VRE layout for optimal resource usage.
- **Content Source Selection:**
  - ✓ Filters out collections unlikely to be contain information sought.
  - ✓ Exploits query-supplied terms and automatically pre-constructed Content Source Descriptors.
- **Query Planning:**
  - ✓ Cost based optimization performed.
  - ✓ Heuristics and space-search.
- **Process Execution:**
  - ✓ Process optimizer selects and allocates appropriate resources to carry out tasks.
- **On-the-Spot processing:**
  - ✓ ResultSet mechanism allows local filtering of large XML chunks of data.
- **Further mechanisms for efficient searches:**
  - ✓ Forward & inverted Indices.
  - ✓ ResultSet transport mechanism to bypass WS-* shortcomings and facilitate paged data exchanges.

More information about D4Science

- Go to the D4Science main website
  www.d4science-project.eu

- Contact the gCube main website
  http://www.gcube-system.org/

Conclusions

- eInfrastructures facilitate the creation of Virtual Research Environments and VRE Ecosystems
- Dynamic content and service provision
- Open access to content w/ open source services
- DRIVER and D4Science strategic projects co-leading the way