DILIGENT: Deploying Virtual Research Environments on-demand

Donatella Castelli
ISTI-CNR
DLs are evolving into “Virtual Research Environments” (Collaboratoria)

- Distributed frameworks for carrying out cooperative activities like “in silico experiments”, data analysis and processing, production of new knowledge using specialised tools
- Largely based on retrieval and access of always updated knowledge from diverse heterogeneous content sources
- The knowledge produced is preserved and made available for other usages inside and outside the VRE
Highly dynamic, created and dismissed on-demand

Based on specialised tools which support the generation of new knowledge
The cost of a dedicated system can be too high for volatile VREs that use many resources.
Outsourcing to the e-Infrastructure

- Shared Resources
- Management and Orchestration
Success factors/challenges

- Infrastructure sustainability
  - Mechanisms for reducing the cost of the infrastructure management

- Supported VREs
  - Flexible and high quality solutions for satisfying the needs of many different applications domains
  - Simple procedures for creating VREs
DILIGENT achievements

- ImpECt: Environmental Monitoring
- ARTE: Education in the Humanities

SAPIR-enabled AV search

DILIGENT Infrastructure

gCube System
Simplifies the infrastructure management

- Resources registration, monitoring, notification,…
- Service deployment, dynamic reallocation, …
- Service composition
Simplifies the construction of a VRE system

- Transparent selection and orchestration of resources by
  - Offering a GUI
  - Abstracting over complexity
  - Abstracting over heterogeneity
VRE generator [cont.]

VRE Generator

Preservation Data kit

gCube Data kit

gCube Mw
Provides flexible search and management functionality
Focus: Search Management

- Most important framework for Information Spaces
- Most important functionality / service in Information Access
VRE Management

From Digital Objects to Content across eInfrastructures

Pasquale Pagano
CNR-ISTI
gCube Monitoring
Continuous analysis of VRE resources status

gCube Keeper
Autonomic management of VRE resources and management of non-critical failures

Rome, 29-30th October 2007
European Information Space: Infrastructures, Services and Applications Workshop

VRE System Administrator
VRE creation, VRE maintenance, and supervision of critical failures

gCube Mw makes the VRE management sustainable
Service Deployment

Rome, 29-30th October 2007

European Information Space: Infrastructures, Services and Applications Workshop

Site 1
- Deploy(services, DHNs)
- DLManagement DHN
- Integrated Feedback Report
- CheckDep(services)
- Deploy(packages)
- Package Repository DHN
- Download(packages)

Site 2
- DIS DHN
- notify()

Site 3
- DHN

Site 4
- DHN

Publish(RI)

Client system
- Browser (UI)
- Deploy
- Deployment Report
- CheckDep(services)
- Deploy(packages)
- Package Repository DHN
- Download(packages)

DILIGENT
- Grid
- External
- gLite

Grid
- DILIGENT
- External
- gLite
ImpECt VRE

From Digital Objects to Content across eInfrastructures
What's behind the scene...

- ES Communities usually operate over widespread geographic scales
  - Scientific collaborations do not take advantage of shared spaces, resources and knowledge
  - Usually very strict time constraints
- Existing infrastructures consist of operational tools and systems which often do not interface with the ones of other institutions
  - Large international initiatives adopt a number of different systems, applications and services that must interface to exchange and process information
  - Evident fragmentation of services
  - Enormous problems of interoperability either at services and data level increased by lack of standards, common agreements and computing/storage resources availability
Data Sources from the Web...

- ~4700 global data set available
  - www.fao.org/geonetwork
  - www.medspiration.org/products
  - www.gmes.info
  - www.eoportal.org
  - www.eea.eu.int

- Environmental data / reports
  - ES Thesaurus ~30000 objects
    - idn.ceos.org
    - eogrid.esrin.esa.int
    - www.medspiration.org/products

- Reference doc Metadata, services
  - www.fao.org/geonetwork
... enriched by a private basket where objects of any type can be accumulated and re-used:

- to fill-in report,
- to create personal collections,
- to populate courses,
- ...
Main Features

- Exploits large infrastructure including EGEE PPS sites
- Accessible via dedicated web portal
- Content produced by a number of data providers (ESA, FAO, EEA, MTS et al.) is organized in collections
- Editable metadata available in different schemas and standards
- Cross-collections and geospatial search
- Content annotation
- On-demand services composition and Grid task submission
- Persistent user area to store reports, query result sets, processes outputs
- Reports composition (periodically revised, kept up to date, published) and document templates definition
- Integration with e-learning platform
The ImpECt VRE has been designed and created on demand by

- Selecting existing collections
- Importing new datasets
- Selecting existing services
- Registering new services
- Orchestrating services by exploiting the workflow framework embedded in the infrastructure
- Deploy everything automatically
- Create automatically the GridSphere portal
Main Objectives

- An open, feature-rich, inherently-distributed Search Engine
  - Composed out of diverse, autonomous, pluggable elements
  - Capturing complex application scenarios combining
    - Information retrieval
    - Data processing

- Maximization of resources placed at the disposal of VRE managers and users
  - Ease of sharing of resources, avoiding mis-utilization and misuse
  - Reduction of cost of ownership and use
Objective: Optimal Utilization of Resources

- Essential for:
  - Maintaining QoS contracts
  - Confronting infrastructure-raised challenges
  - Attracting resources to the Grid

- Special challenges:
  - Uncontrolled and dynamic environment
  - High-dimensional search space
  - Multi-facet quality metrics
  - Heterogeneity
Search Management: orchestration of search services

Operation highlights:

- Planning & Optimization
- Distributed Information Retrieval
- Incremental result delivery
Retrieval of Distributed Information

Distributed Retrieval of Information
• System diversity
  • Internal, registered/indexed by the system
  • External, Google, JDBC data sources, ISIS/OSIRIS system

• Data diversity
  • Structured and semi-structured (xml)
  • Images
  • Geospatial and temporal
  • Potentially thematically focused

• Processing diversity
  • Metadata structures
  • Querying cost
  • Ranking estimation
THE CHALLENGE

- Characterizing and indexing a diversity of sources
- Selecting the appropriate sources
- Fusing/Merging the results in meaningful lists
Indexing for Content Based Search

Query

Portal

Feature Extraction

Index Mgt

Metadata & Content Mgt

Content & Metadata

Extract features

Build Index

Query Index

MD

Feed

Access metadata & create ResultSet

Feature Extraction Service

Feature Index

Present results
- Numerous Search services, for info retrieval & processing
  - Structured data and XML processing (scanners, sorters, joiners, filterers, transformers, retrievers)
  - Lookups (indices, FT indices, XML indices, Geo indices)
  - Content-based searches
  - External source probes
  - Fusion / Merging of results
- Query language (internal) for interfacing
- Workflow language (BPEL) for execution
- Data transport mechanism (ResultSet) for communication
project by 'title', 'description', 'subject'
on (keeptop 20
on (sort ASC by 'DocID'
on (merge
(fieldedsearch by 'title' contains '*woman' in 'ENGLISH'
on 'CollectionOfMedicalImages'
as 'dc'))
and (fieldedsearch by 'description' contains '*term*' in 'ENGLISH'
on 'CollectionOfMedicalBooks'
as 'dc'))
)
)
)
)

Queries & Workflows: It can get complex...

project by 'title', 'date' on
(sort ASC by 'DocID' on
(merge on
  //MAP REPORTS
  keeptop 8 on
  (sort ASC by 'RankID' on
   (join inner by 'DocID' on
    (fulltextsearch by 'Mediterranean' in 'ENGLISH' on 'd369b3e0-fa4c-11db-a297-9c01d805f283')
    and
    (fulltextsearch by 'Environmental' in 'ENGLISH' on 'd369b3e0-fa4c-11db-a297-9c01d805f283'))))))

keeptop 8 on (sort ASC by 'RankID' on (join inner by 'DocID' on (fulltextsearch by 'Mediterranean' in 'ENGLISH' on 'd369b3e0-fa4c-11db-a297-9c01d805f283') and (fulltextsearch by 'Environmental' in 'ENGLISH' on 'd369b3e0-fa4c-11db-a297-9c01d805f283')))

// EEA reports
keeptop 8 on
  (sort ASC by 'RankID' on
   (fieldedsearch by 'date' contains "1999" on
    (join inner by 'DocID' on
     (fulltextsearch by 'air pollution' in 'ENGLISH' on '25ad3c50-fa41-11db-a270-9c01d805f283')
      and
      (fulltextsearch by 'european' in 'ENGLISH' on '25ad3c50-fa41-11db-a270-9c01d805f283')))
Optimal Utilization of Resources

- **Pre-query optimization:**
  - Monitoring and adaptation of VRE layout for optimal resource use

- **Content Source Selection:**
  - Filtering of collections unlikely to contain useful data
  - Query terms and automatically pre-constructed Content Source Descriptors

- **Query Planning:**
  - Cost based optimization
  - Heuristics and space-search

- **Process Execution:**
  - Process optimization selects and allocates appropriate resource for tasks

- **On-The-Spot processing:**
  - ResultSet mechanism to allow local filtering of large XML chunks of data

- **Further mechanisms to facilitate efficient searches:**
  - Indices
  - ResultSet transport mechanism
From Digital Objects to Content across eInfrastructures

from theory ...

... to reality