DILIGENT: Deploying Virtual Research Environments on-demand

Donatella Castelli, Pasquale Pagano
ISTI-CNR
Yannis Ioannidis
Univ. of Athens
Outline

- Motivations & overview
- Achievements
  - DL related services
  - DILIGENT Infrastructure
  - ImpECt application
- D4Science
Motivations – from DLs to VREs

- 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
VRE system

VRE

VRE System

Diligent

Dedicated Resources
Content Sources
Services
Computing & storage elements
Management and Orchestration
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

e-Infrastructure
Success factors/challenges

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

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

SAPIR-enabled AV search

ImpECt
Environmental Monitoring

ARTE
Education in the Humanites

gCube System

DILIGENT Infrastructure

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

- Resources registration, monitoring, notification,…
- Service deployment, dynamic reallocation, …
- Service composition
gCube Middleware [cont.]

- Preservation Data kit
- gCube Data kit
- gCube Mw
Simplifies the construction of a VRE system

- Transparent selection and orchestration of resources by
  - Offering a GUI
  - Abstracting over complexity
  - Abstracting over heterogeneity
Provides flexible search and management functionality
Focus: Search Management

Most important framework for Information Spaces

Most important functionality / service in Information Access
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
Selecting Sources and Fusing 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 (keep top 20
on (sort ASC by 'DocID'
on (merge
(filedsearch by 'title' contains '*woman'
in 'ENGLISH'
on 'CollectionOfMedicalImages'
as 'dc')
and (fieldedsearch by 'description' contains '*term*'
in 'ENGLISH'
on 'CollectionOfMedicalBooks'
as 'dc'))
)
)

Optimization
Complex Cost Calculation
Profiling / Monitoring
Resource selection "hinting"
Domain specific planning

Parallelization

Active Planning
project by 'title', 'date' on
(sort ASC by 'DocID' on
(merge on
//MAP REPORTS
keep top 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')))
keep top 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
keep top 8 on
(sort ASC by 'RankID' on
(fieldedsearch by 'date' contains "1999" on
(join inner by 'DocID' on
(fulltextsearch by 'air polution' 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
Next step: DILIGENT for Science

- Provide and operate a production D4Science e-Infrastructure
- Consolidate and extend gCube
- Built VREs serving Environmental Monitoring and Fishery Resources Management domains
Main technological challenges

- Provide and operate a production D4Science e-Infrastructure
  Define the operational procedures for sites (sites include content and service sites)

- Consolidate and extend gCube
  Extend the Data Kit to deal with very large and heterogenous content sources (e.g. textual repositories, satellite images, statistical databases) and other content-related resources (e.g. gazetters, ontologies, thesauri)

- Build VREs serving Environmental Monitoring and Fishery Resources Management domains
  Serve the needs of a multitude of researchers and decision-makers from many disciplines (biologists, climatologists, GIS experts, socio-economists, fishery managers, etc.) operating with many different tools
http://www.diligentproject.org

http://www.d4science.org/
Thank you!  
Questions?
Additional Slides
An **application framework** for the **development** of services that can be outsourced to a grid-enabled infrastructure

An advanced **container** for the **hosting** of WS on the grid

A **runtime environment** for the

- provision of information about shared resources
- management of services and applications
- execution of VRE build-in services: content and metadata management; indexing, selection, fusion, extraction, description, annotation, transformation, and presentation of content
VREs: new requirements

Persistent and consolidated
e.g. serving a team of individuals in addressing the mission of an institution

Focus on publication
e.g. supporting the publishing and archival of content

Highly dynamic, created and dismissed on-demand
e.g. supporting the activities of a project addressing a specific challenge

Analysis and production of new knowledge
e.g. serving a research team which produces new results through complex analysis and simulation