DILIGENT:
Deploying Virtual Research Environments on-demand

Diligent

From Digital Objects
to Content across
eInfrastructures

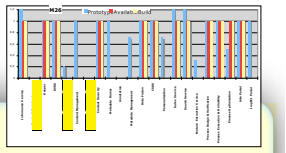
Donatella Castelli
ISTI-CNR

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

VREs trend

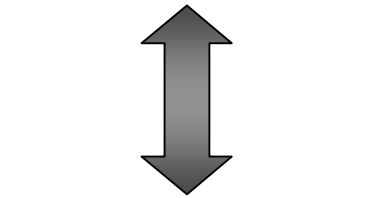
Highly dynamic, created and dismissed on-demand



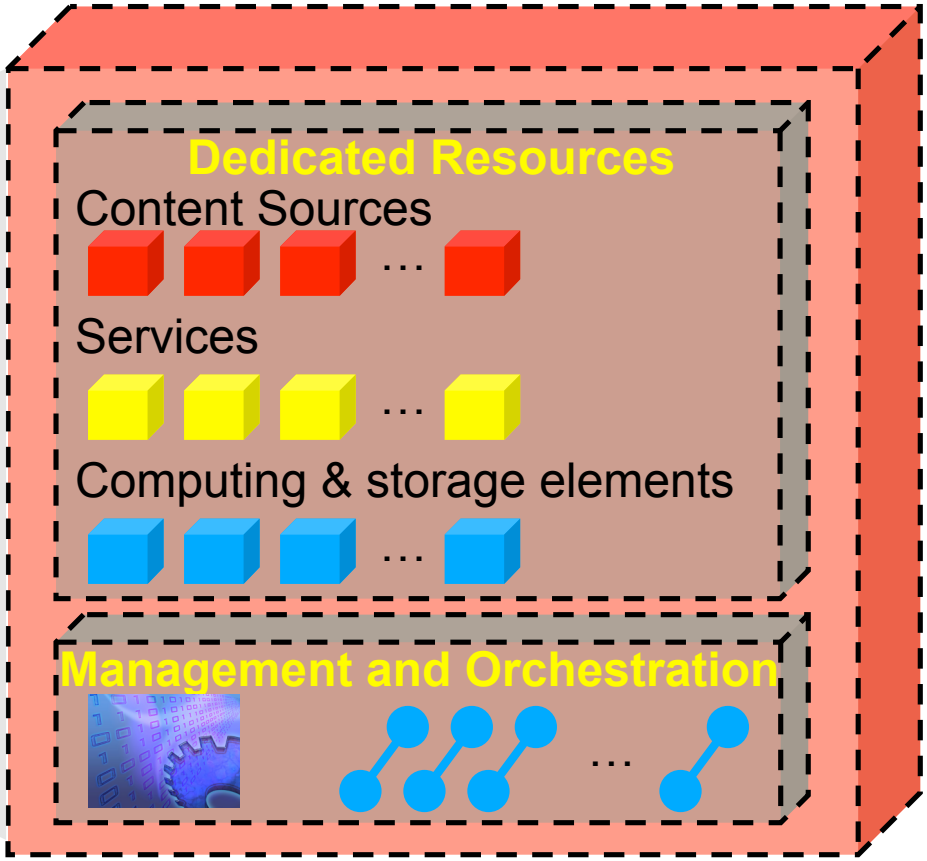
Based on specialised tools which support the generation of new knowledge

VRE system

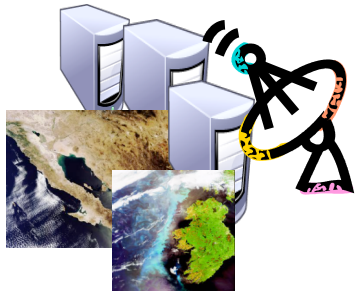
VRE



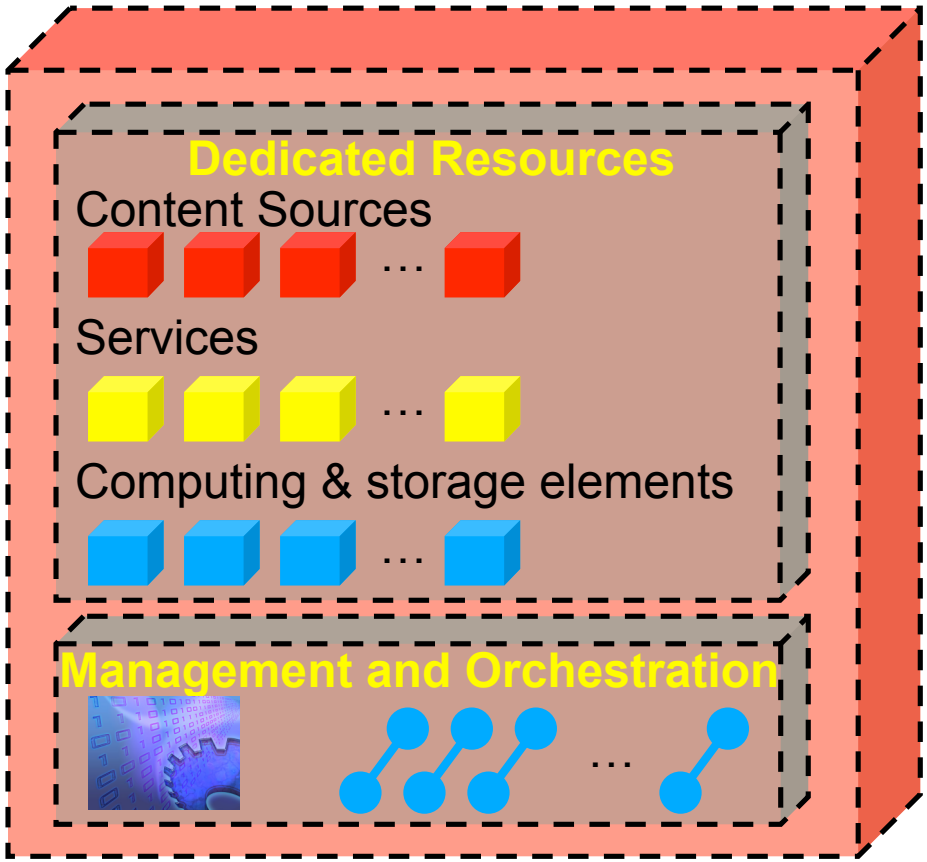
VRE System



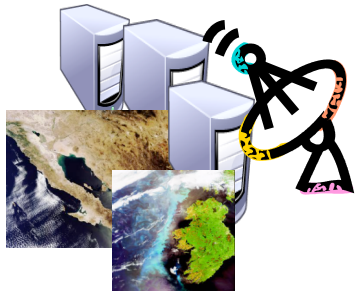
1-to-1 model: sustainability

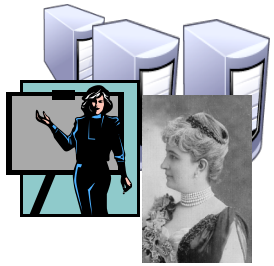


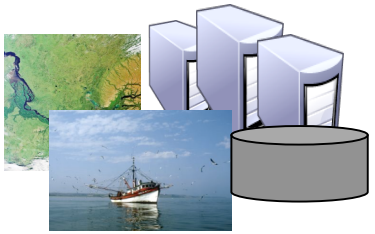
- The cost of a dedicated system can be too high for volatile VREs that use many resources

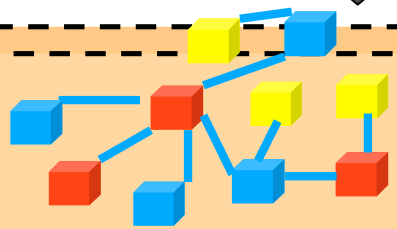


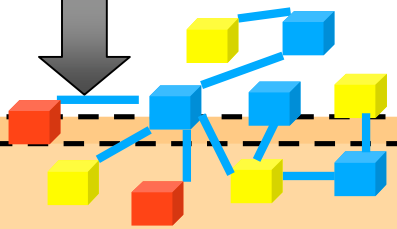
Outsourcing to the e-Infastructure

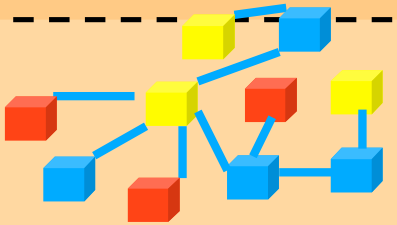




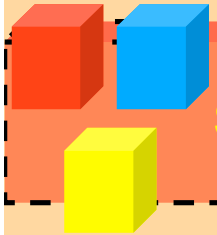








e-Infrastructure



Shared Resources

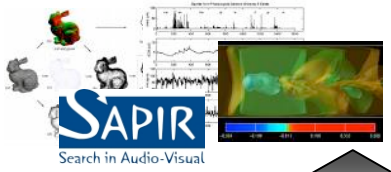
Management and Orchestration

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

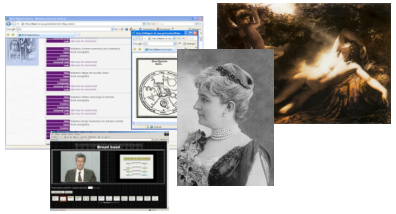
DILIGENT achievements

SAPIR-enabled AV search



ImpECt Environmental Monitoring

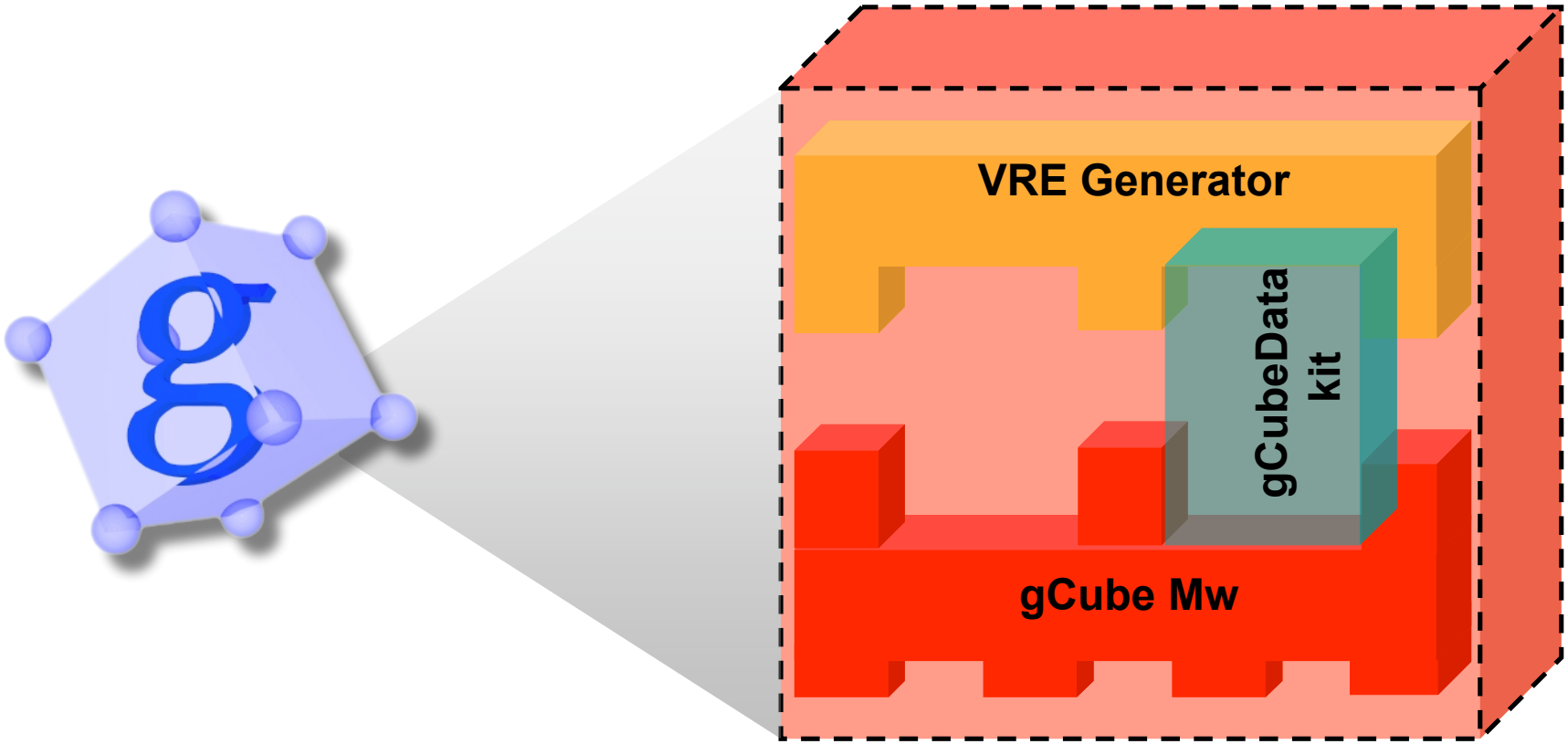
ARTE Education in the Humanites

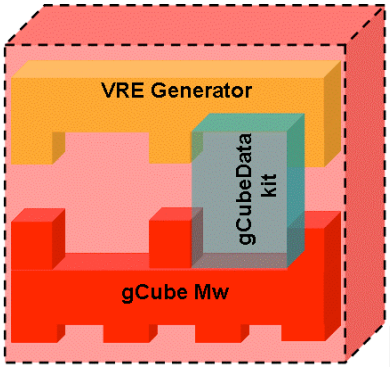


DILIGENT Infrastructure

gCube System

gCube overview



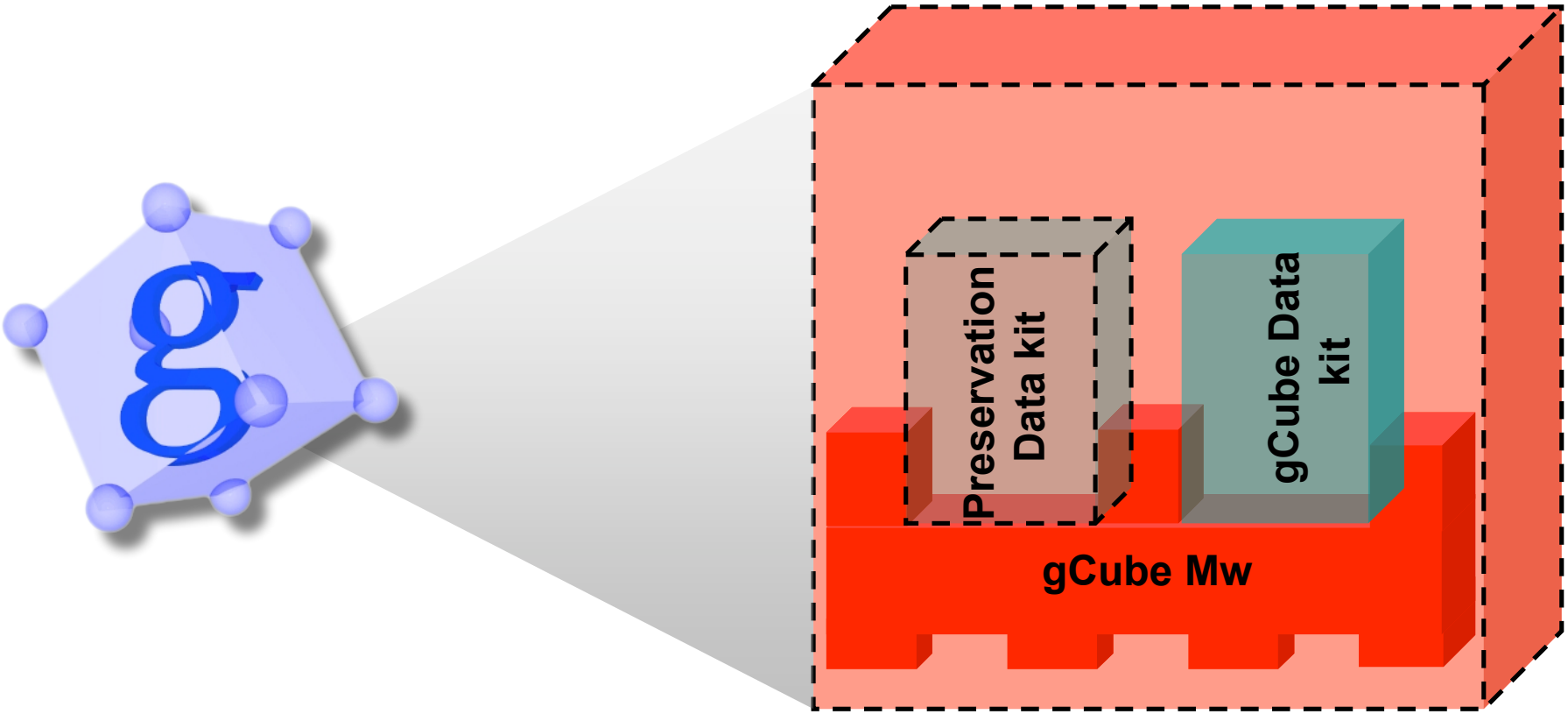


Simplifies the infrastructure management

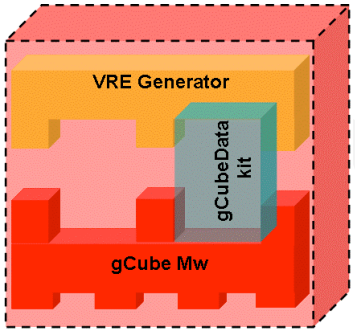
- Resources registration, monitoring, notification,…
- Service deployment, dynamic reallocation, …
- Service composition

Resource | Service | Content Source | Comp&Storage

gCube Middleware [cont.]



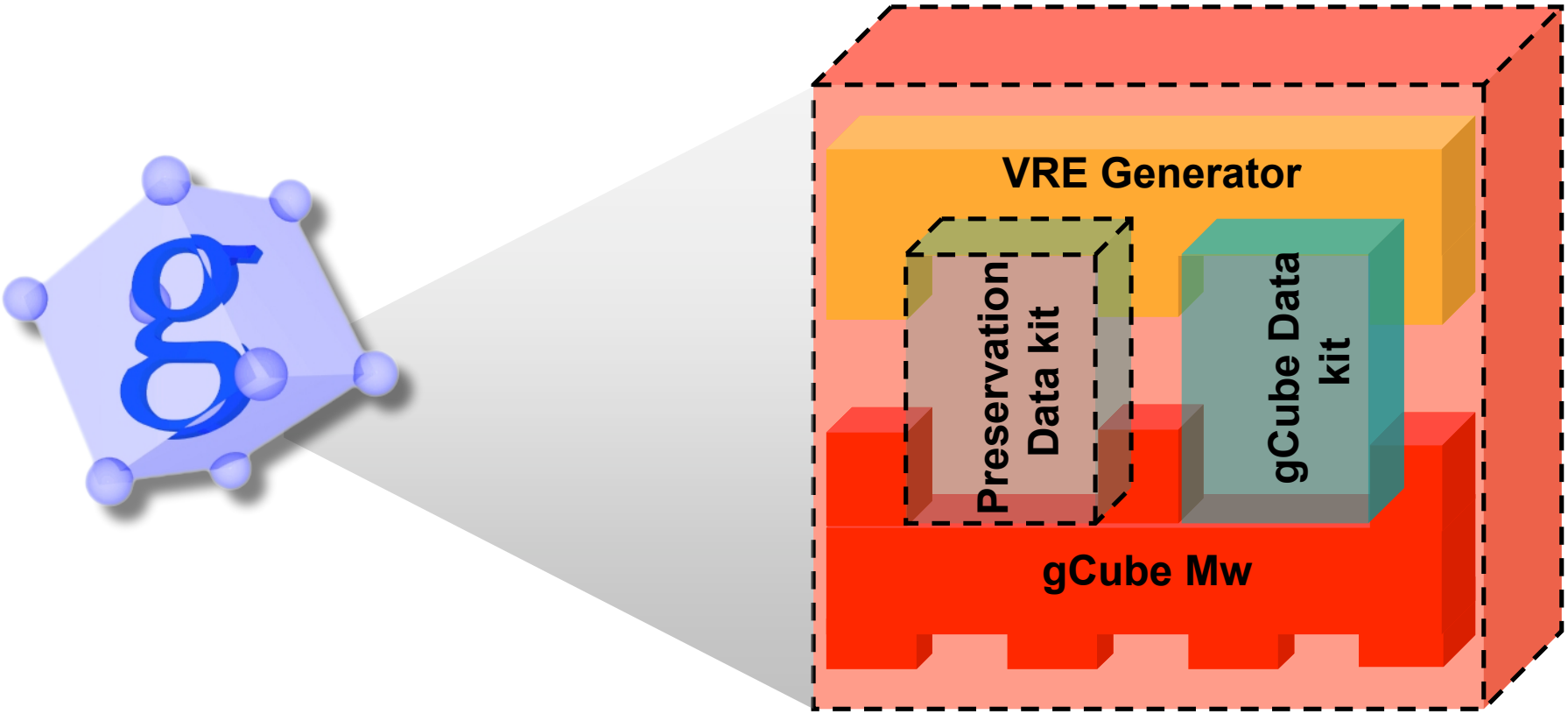
VRE generator



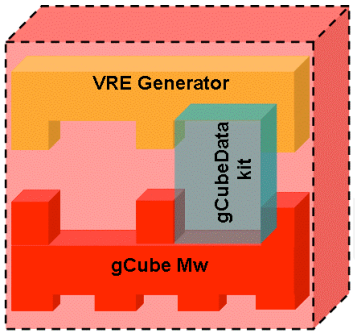
Simplifies the construction of a VRE system

- Transparent selection and orchestration of resources by
  - Offering a GUI
  - Abstracting over complexity
  - Abstracting over heterogeneity

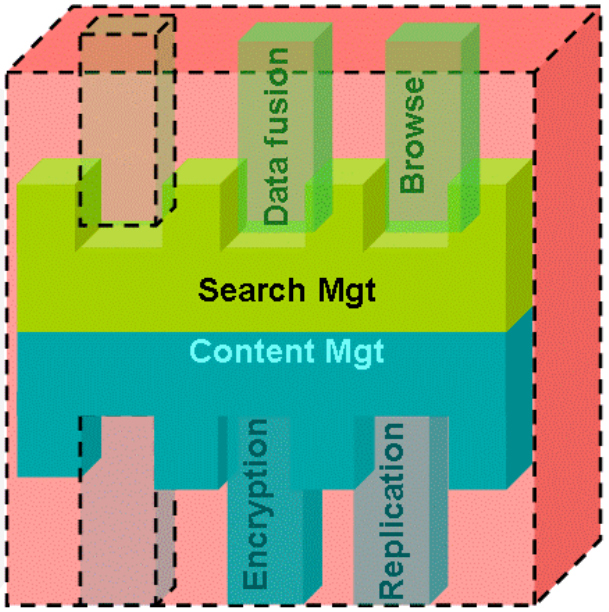
VRE generator [cont.]



gCube Data Kit

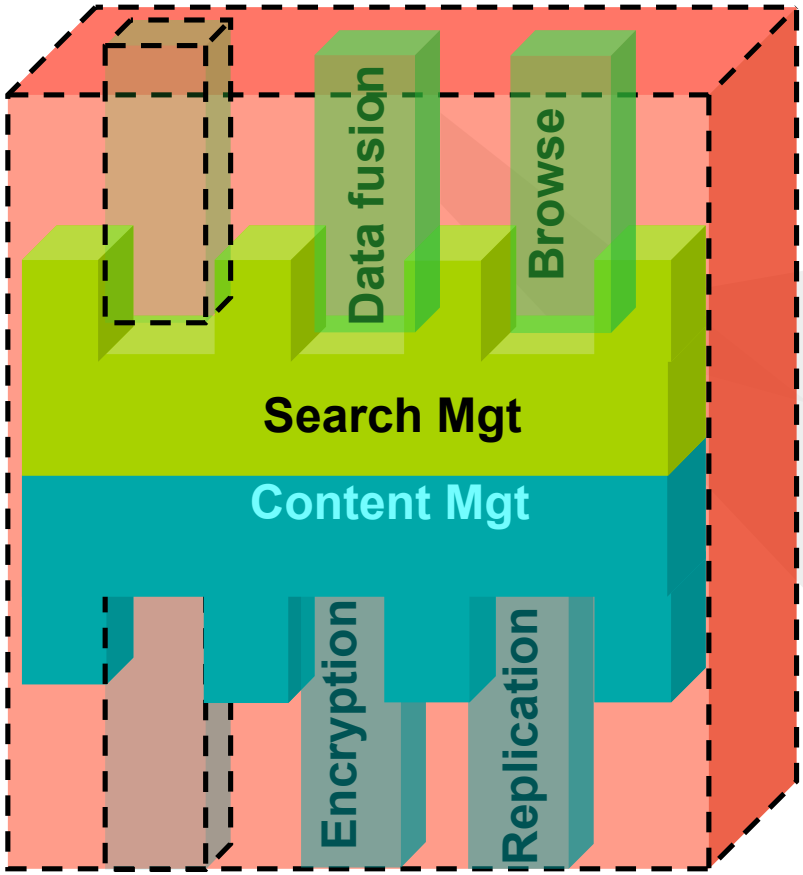


Provides flexible search and management functionality



Data Fusion	Browse	Source sel.	Feature extr.

Focus: Search Management



Most important framework for Information Spaces

Most important functionality / service in Information Access

VRE Management

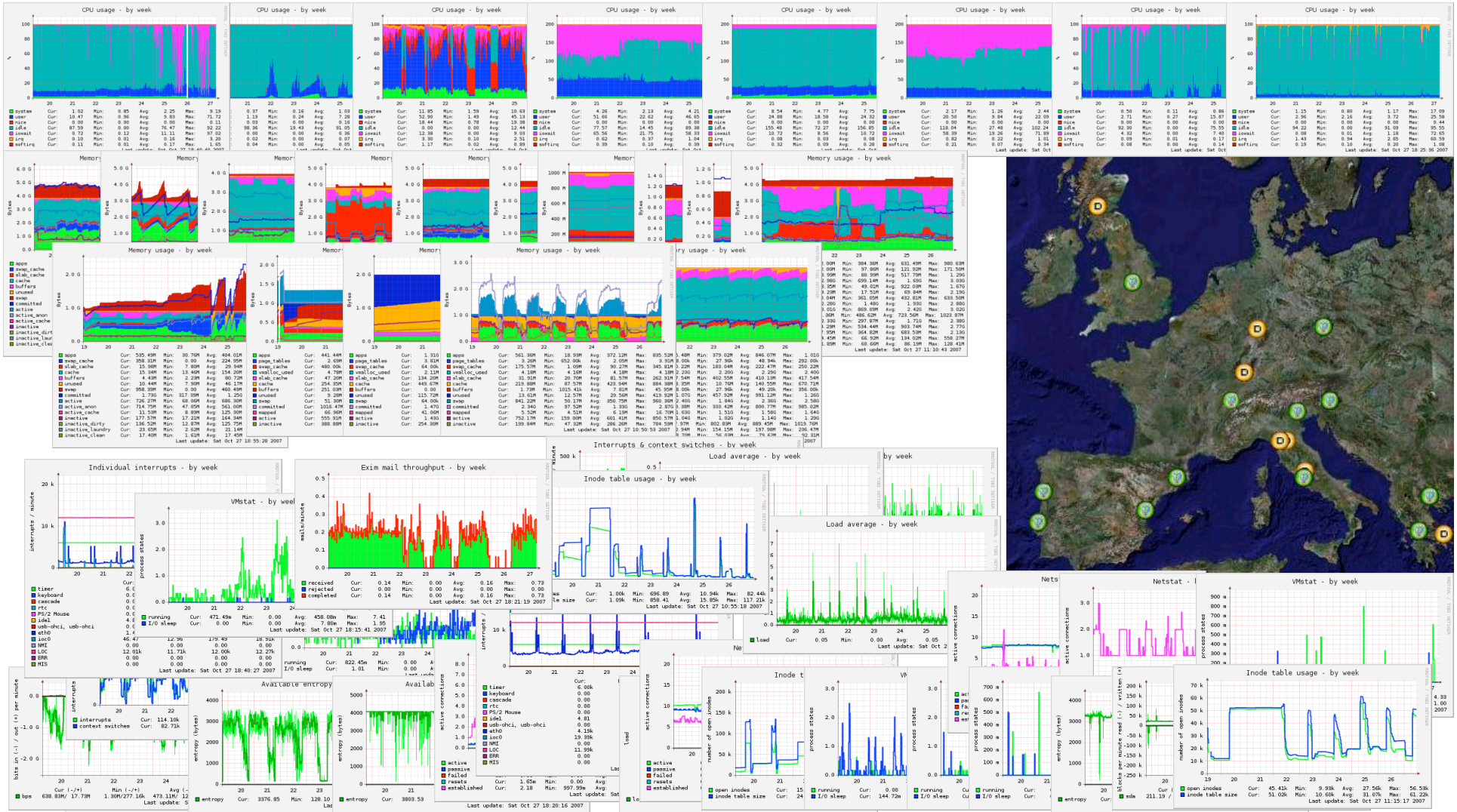
Diligent

From Digital Objects
to Content across
eInfrastructures

Pasquale Pagano
CNR-ISTI

Diligent

Monitoring



gCube Middleware

gCube Monitoring

Continuous analysis of VRE resources status

gCube Keeper

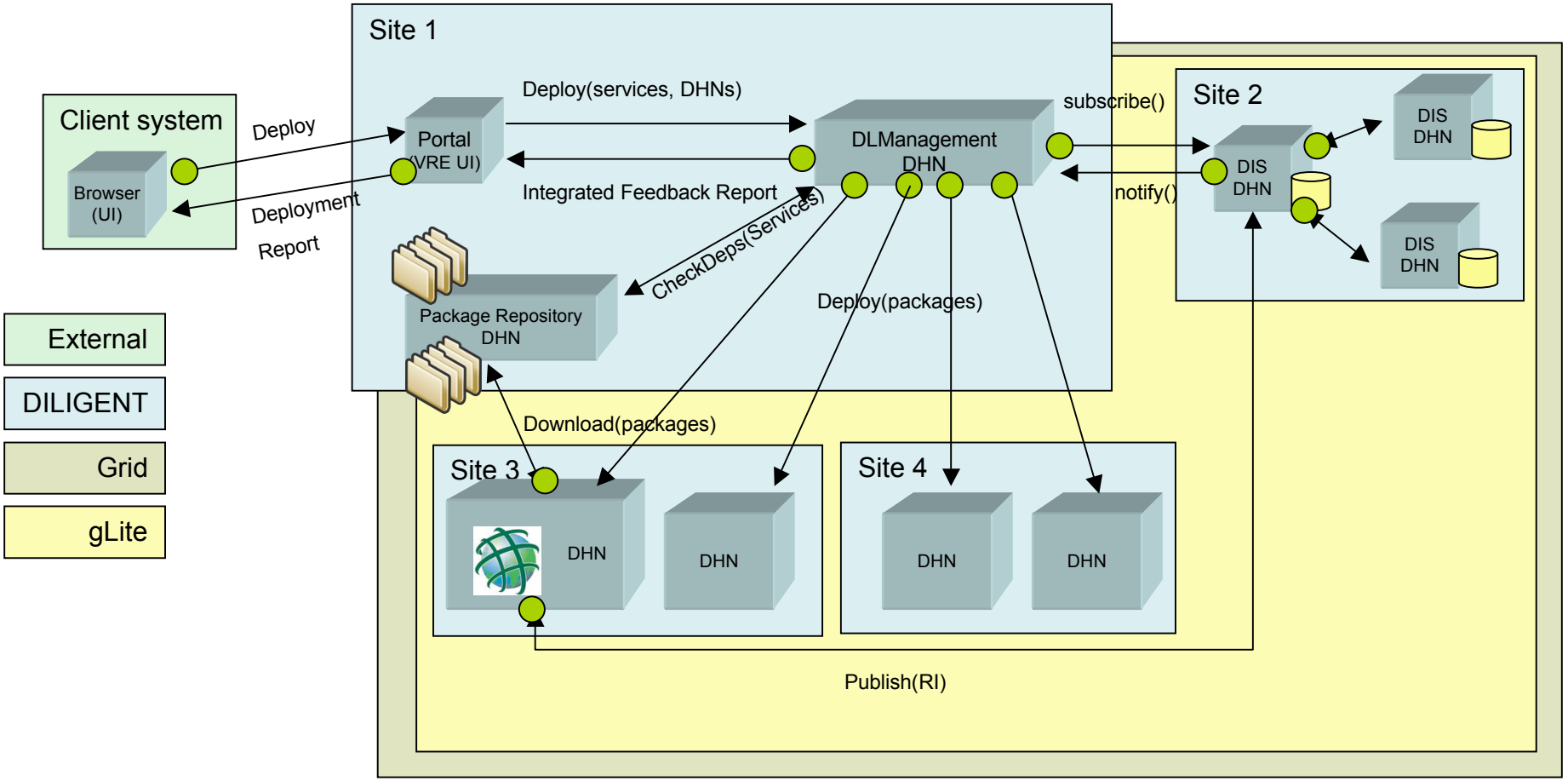
Autonomic management of VRE resources and management of non-critical failures

gCube Mw makes the VRE management sustainable

VRE System Administrator

VRE creation, VRE maintenance, and supervision of critical failures

Service Deployment



ImpECt VRE

Diligent

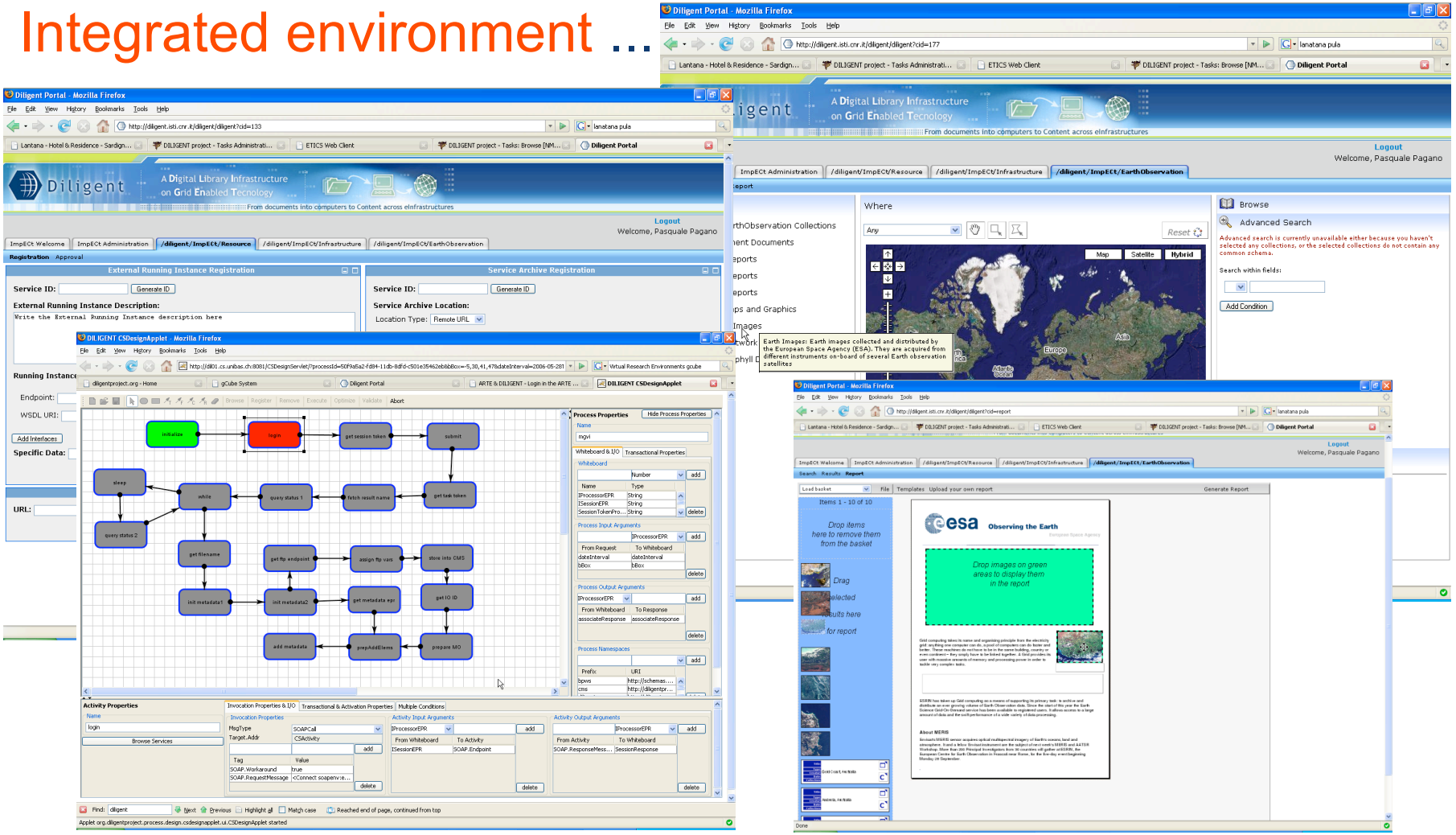
From Digital Objects
to Content across
eInfrastructures

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...

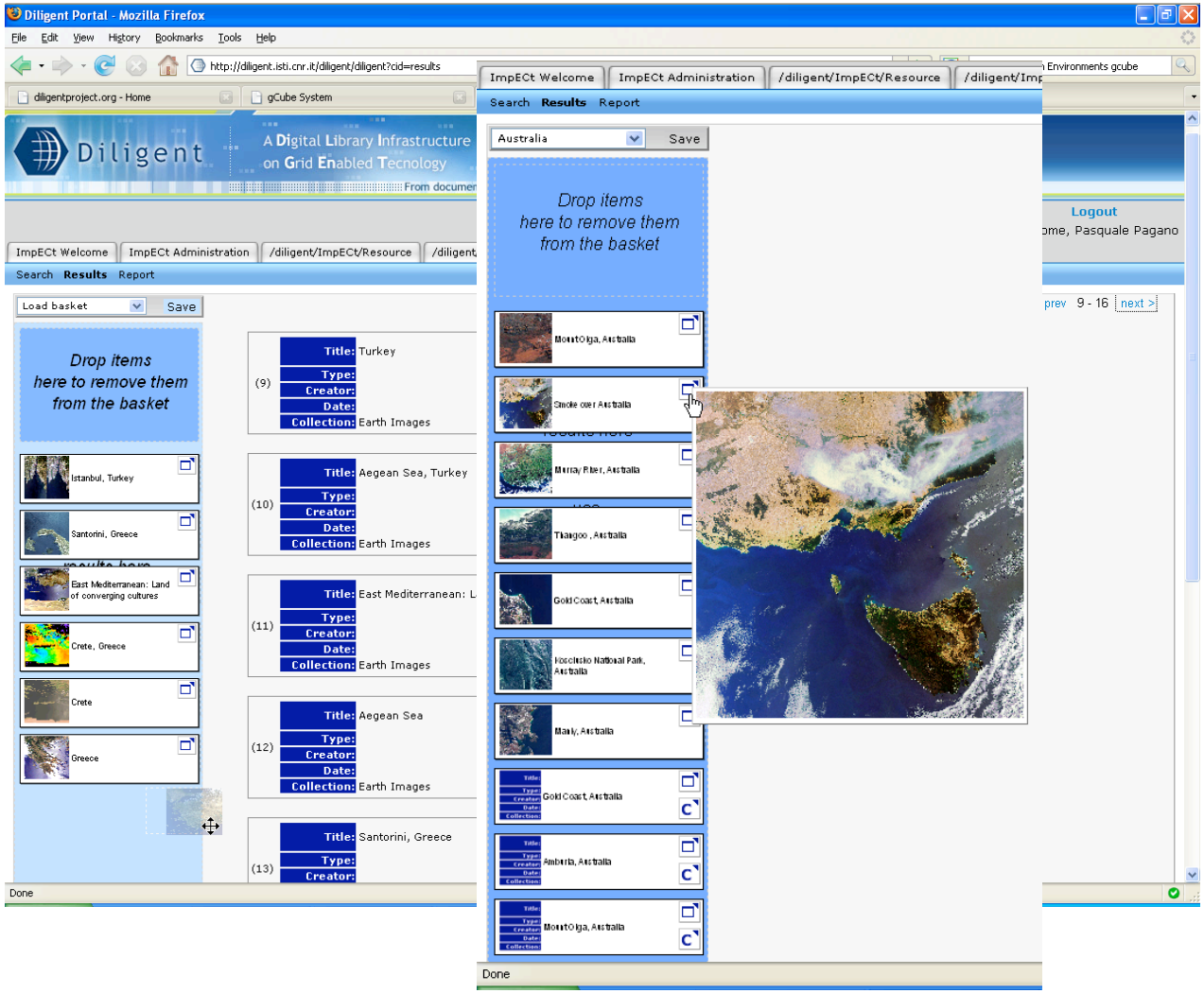
Integrated environment ...



ImpECt : private workspace

… 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

ImpECt VRE creation

- 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 Kit

- Search Management: orchestration of search services
- Operation highlights:
  - Planning & Optimization
  - Distributed Information Retrieval
  - Incremental result delivery

Retrieval of Distributed Information

Distributed Retrieval of Information

Distribution #1: Information Sources

- 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

Distribution #1: Information Sources

- THE CHALLENGE
  - Characterizing and indexing a diversity of sources
  - Selecting the appropriate sources
  - Fusing/Merging the results in meaningful lists

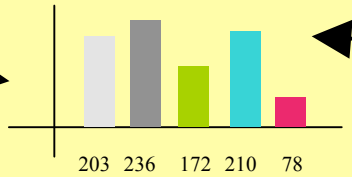
Diligent

Indexing for Content Based Search

Portal | Feature Extraction | Index Mgt | Metadata & Content Mgt | Content & Metadata

Query

Extract features



Build Index

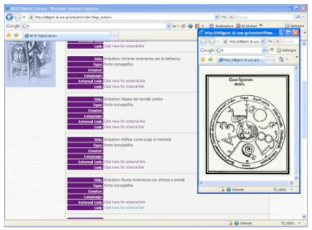
Index

Query Index

MD

Feed

Present results

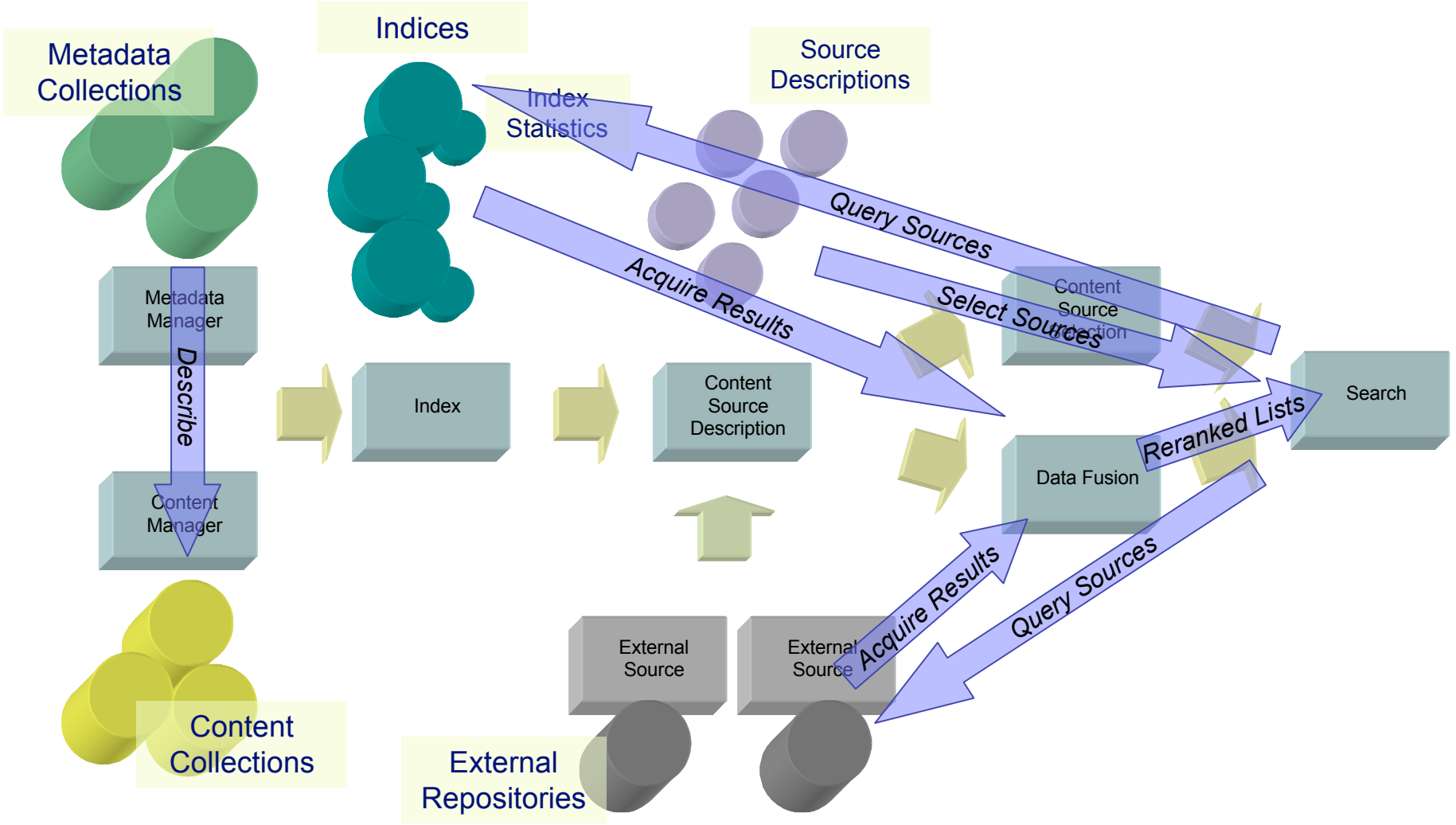


Access metadata & create ResultSet

Feature Extraction Service

Feature Index

Selecting Sources and Fusing Results



Distribution #2: Information Retrieval

- 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

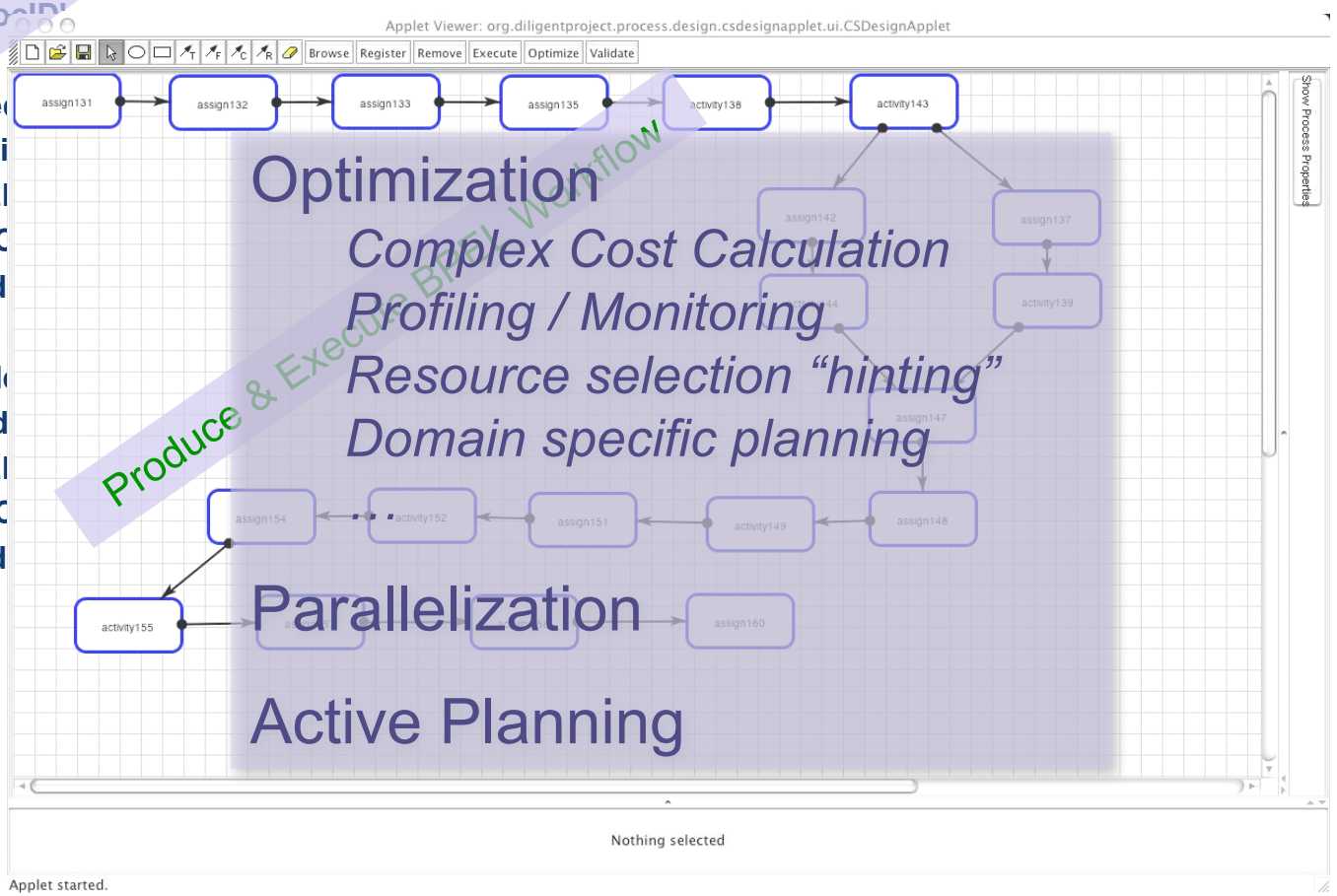
Query and Workflow Management

```
project by 'title', 'description', 'subject'
    on (keeptop 20
        on (sort ASC by 'DocID'
            on (merge
                on (fielde
                    by 'ti
                    in 'El
                    on 'C
                    as 'd

                and (fielde
                    by 'd
                    in 'El
                    on 'C
                    as 'd
                )
            )
        )
```

Query

Produce & Execute BPEL Workflow



Optimization

Complex Cost Calculation

Profiling / Monitoring

Resource selection "hinting"

Domain specific planning

Parallelization

Active Planning

Queries & Workflows: It can get complex...

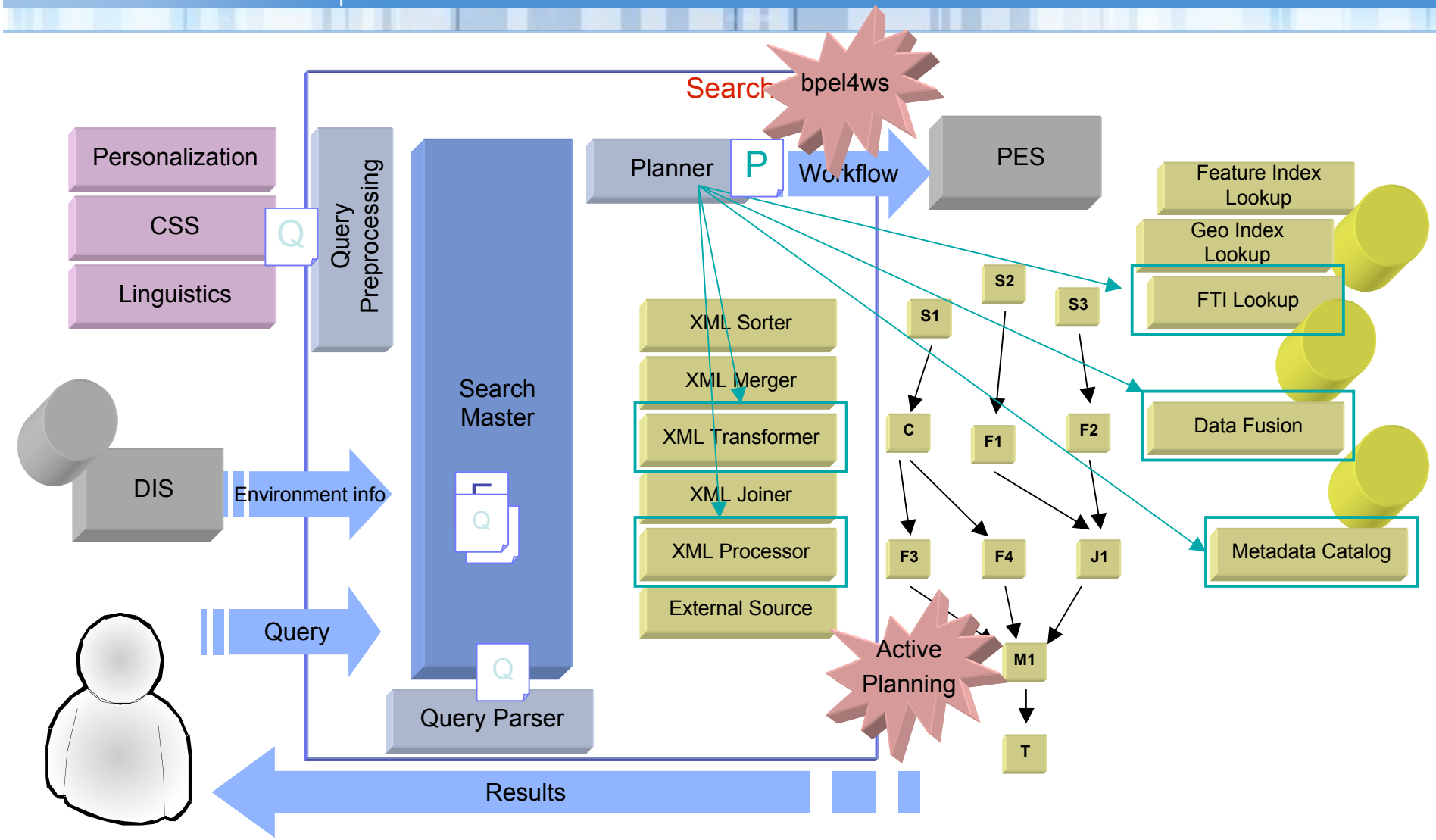
```
project by 'title', 'date' on
    (sort ASC by 'DocID' on
        (merge on
            //MAP REPORTS
            keeptop 8 on
                (sort ASC by 'RankID' on
                    (join inner by 'DocID' on
                        (fulltextsearch by 'Mediterranean' in 'ENGLISH' on 'd369b3e0-fa4c-11db-a297-9c01d805f283')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

Information Retrieval: How it Works



Diligent

From Digital Objects
to Content across
eInfrastructures

from theory ...

... to reality