

# **DILIGENT:** **Deploying Virtual Research Environments on-demand**



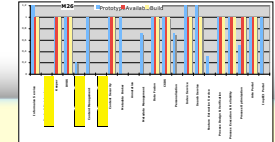
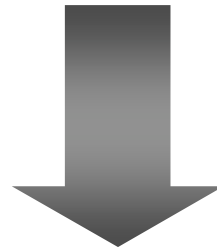
Diligent

From Digital Objects  
to Content across  
eInfrastructures

Donatella Castelli  
ISTI-CNR



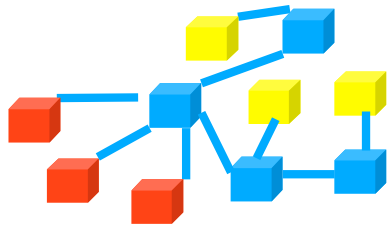
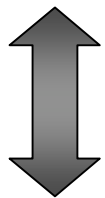
- DLs are evolving into “Virtual Research Environments” (Collaboratoria)
  - Distributed frameworks for carrying out **cooperative activities** like “in silico experiments”, data analysis and processing, production of new knowledge using specialised tools
  - Largely based on **retrieval and access** 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



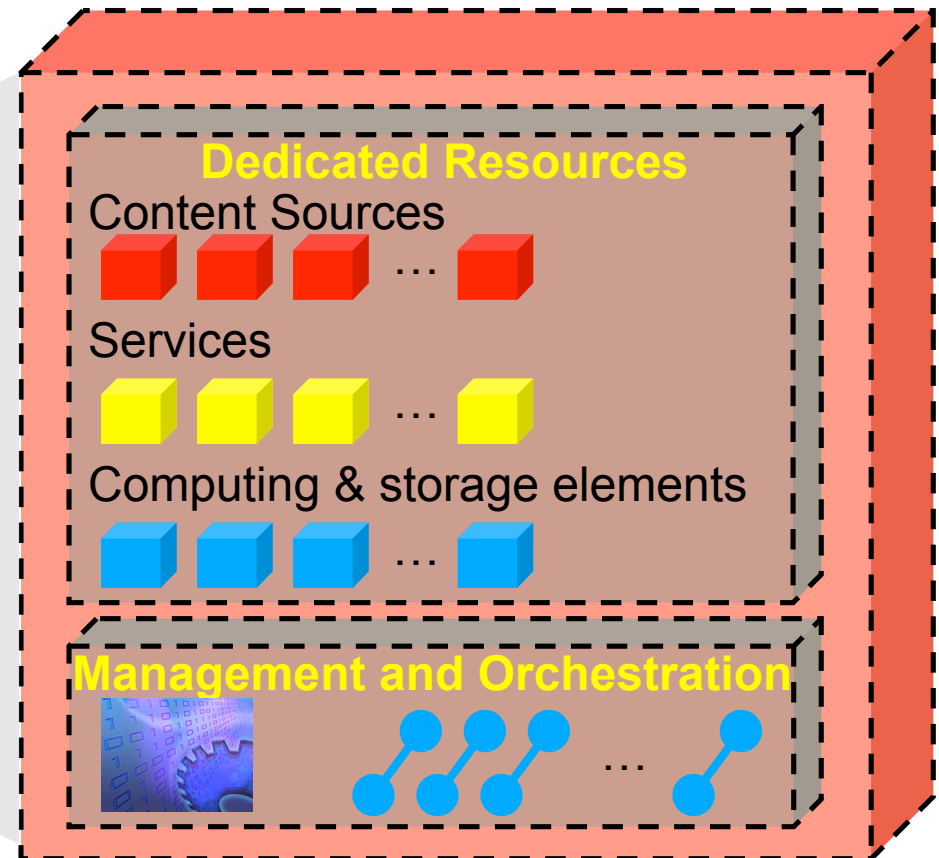
**Highly dynamic, created and dismissed on-demand**

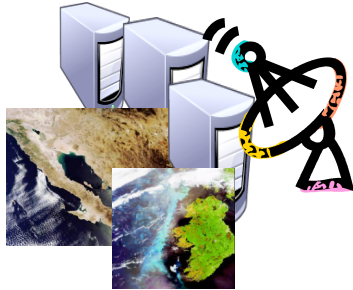
**Based on specialised tools which support the generation of new knowledge**



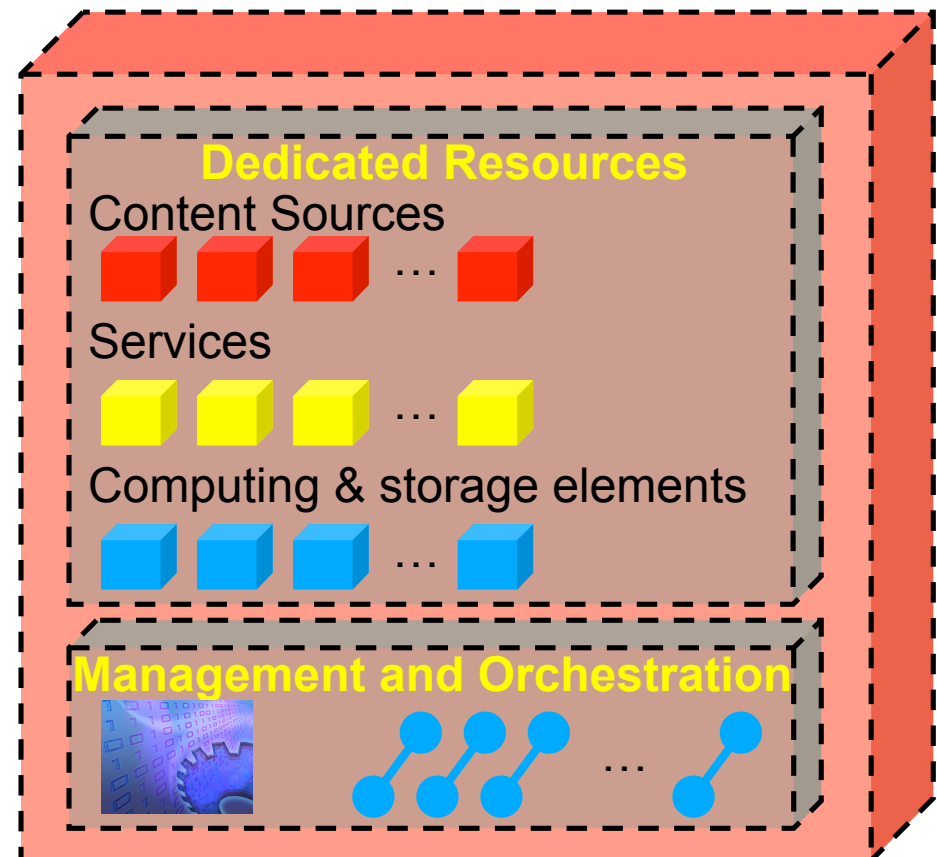


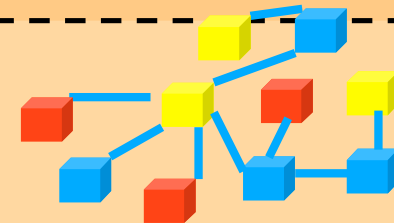
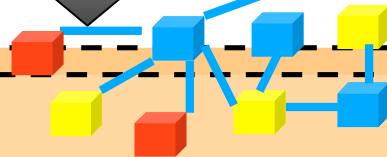
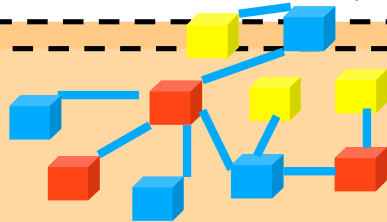
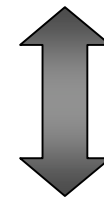
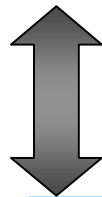
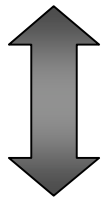
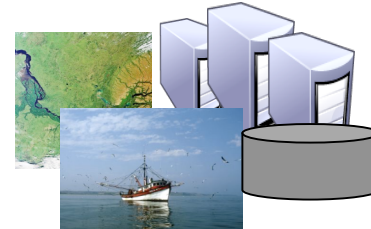
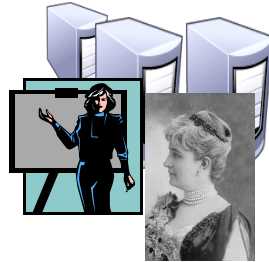
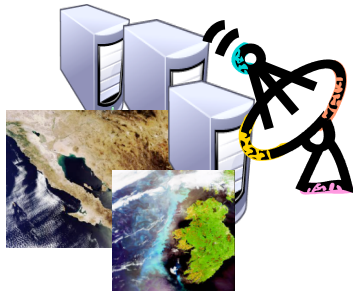
**VRE System**





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





**e-Infrastructure**

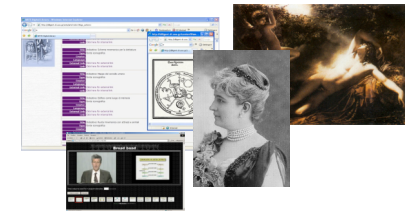


- **Infrastructure sustainability**
  - Mechanisms for reducing the cost of the infrastructure mng
- **Supported VREs**
  - Flexible and high quality solutions for satisfying the needs of many different applications domains
  - Simple procedures for creating VREs

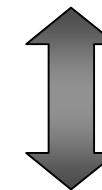
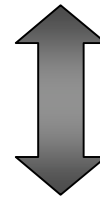
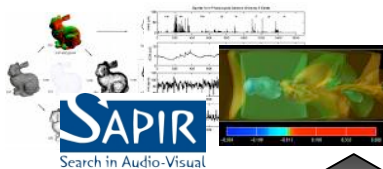
## ImpEct Environmental Monitoring



## ARTE Education in the Humanities



## SAPIR-enabled AV search

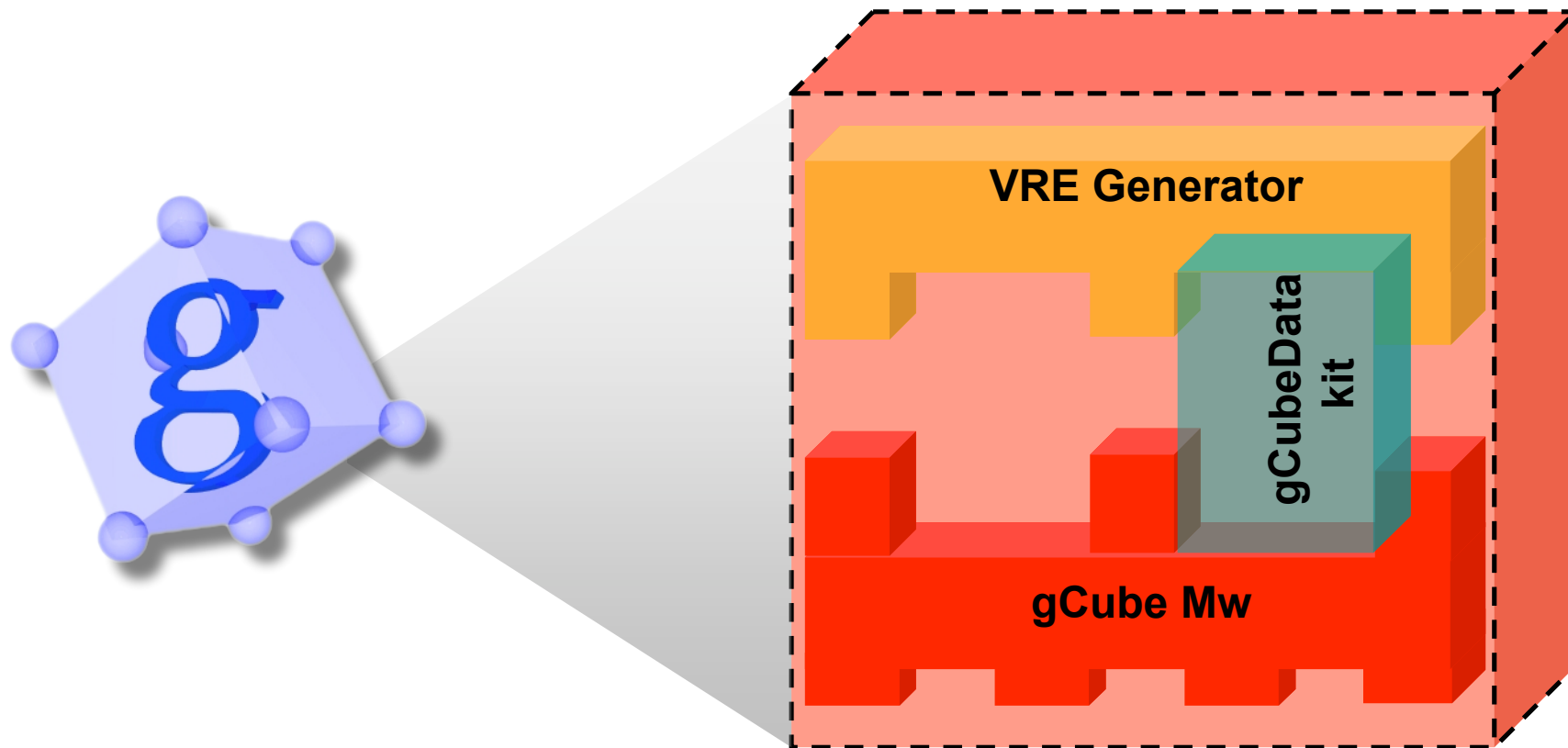


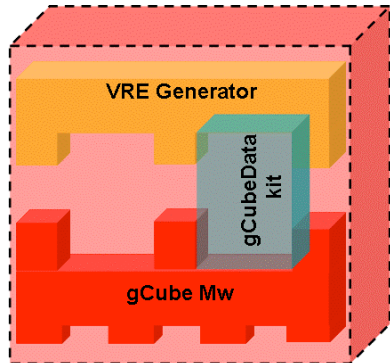
DILIGENT  
Infrastructure



gCube System







## Simplifies the infrastructure management

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



Resource

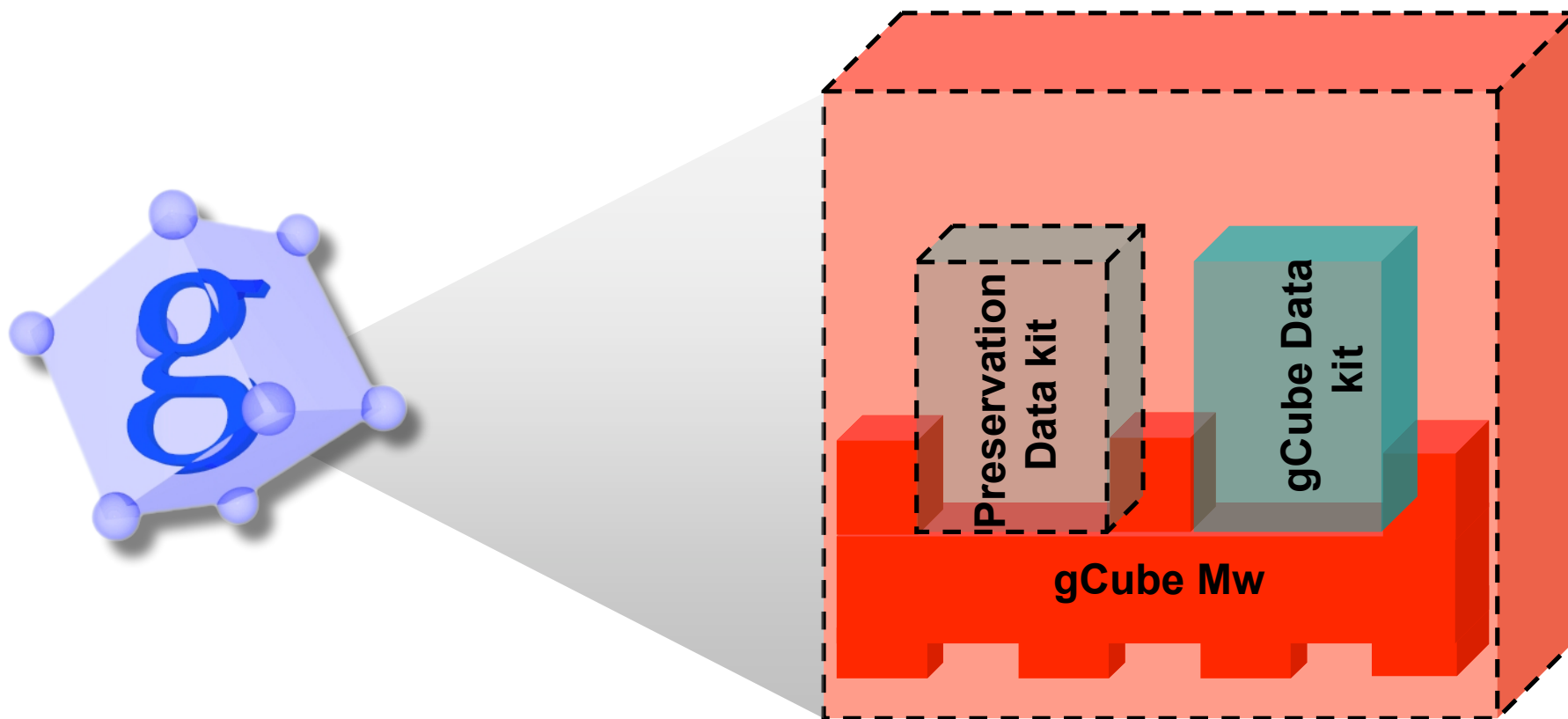


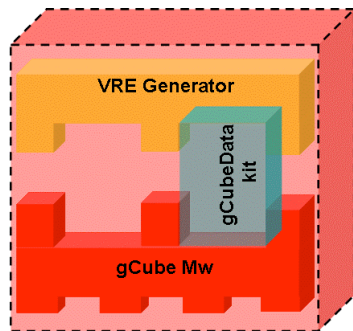
Service

Content Source



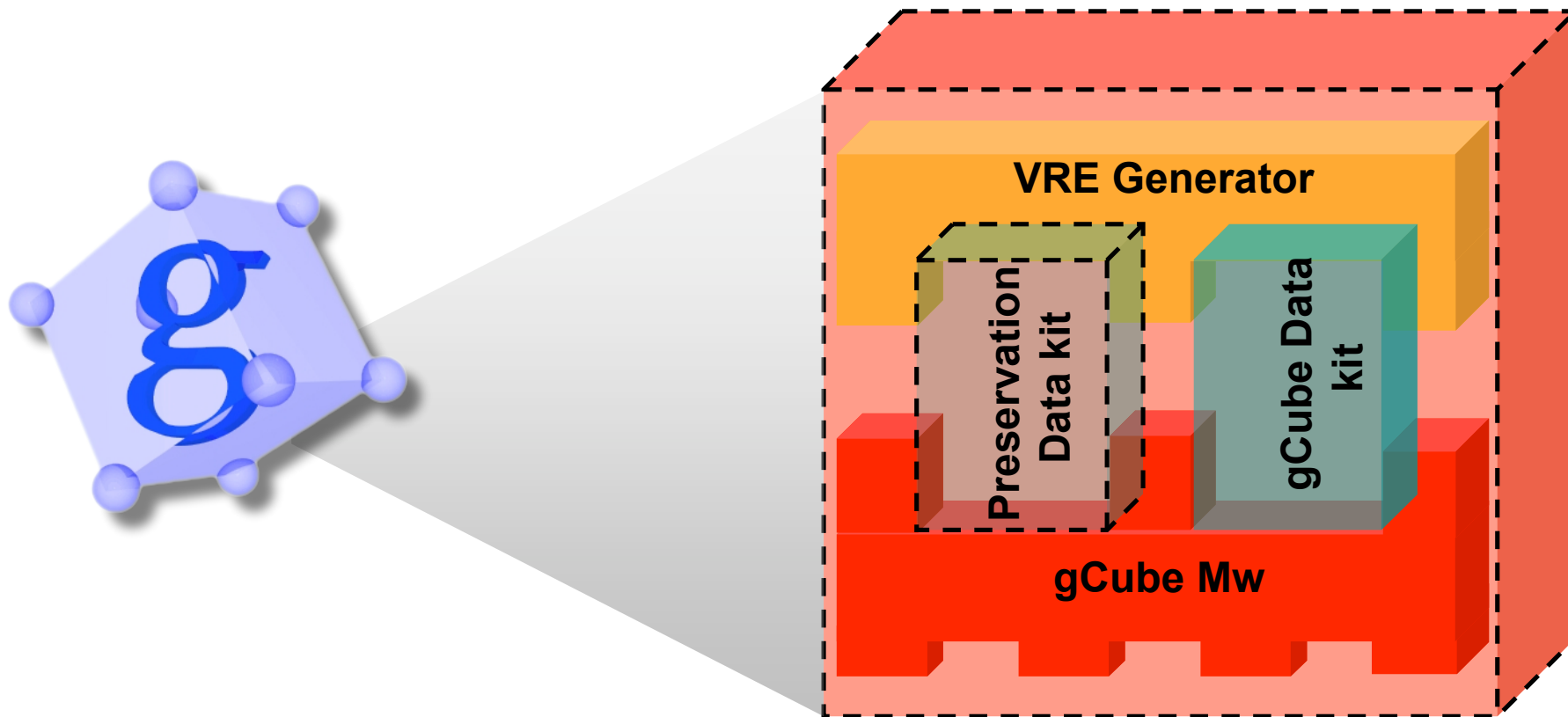
Comp&Storage

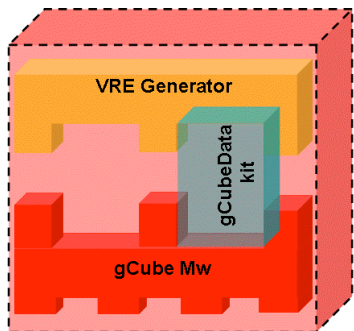




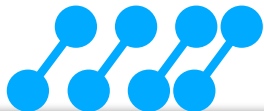
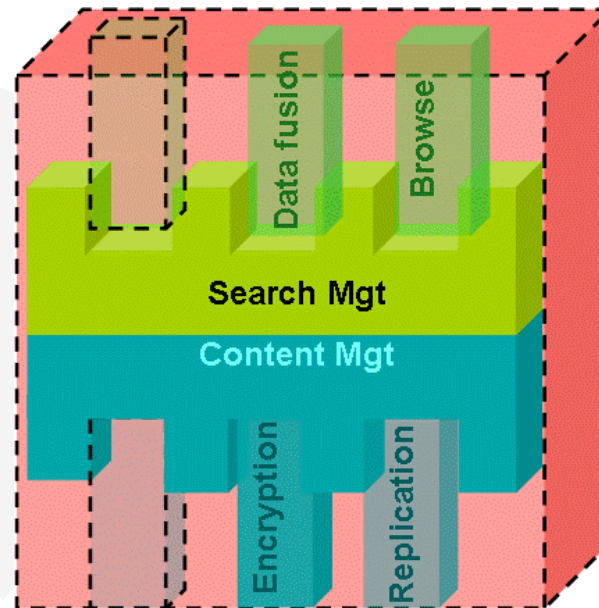
## Simplifies the construction of a VRE system

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





Provides flexible search and management functionality



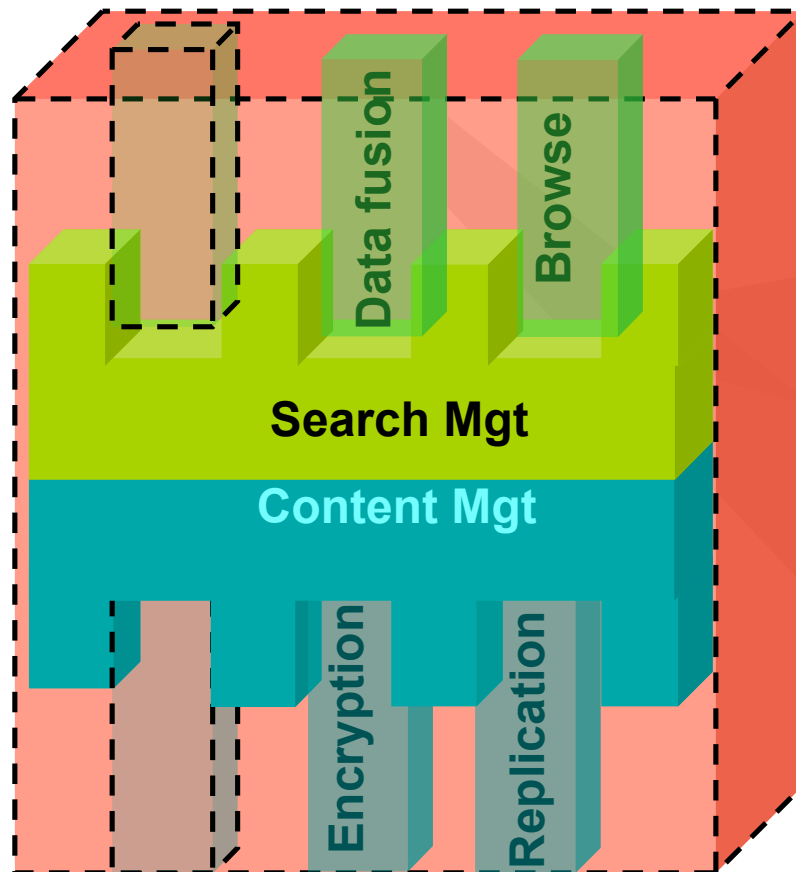
Data Fusion

Browse

Source sel.

Feature extr.





Most important framework for Information Spaces

Most important functionality / service in Information Access

# VRE Management

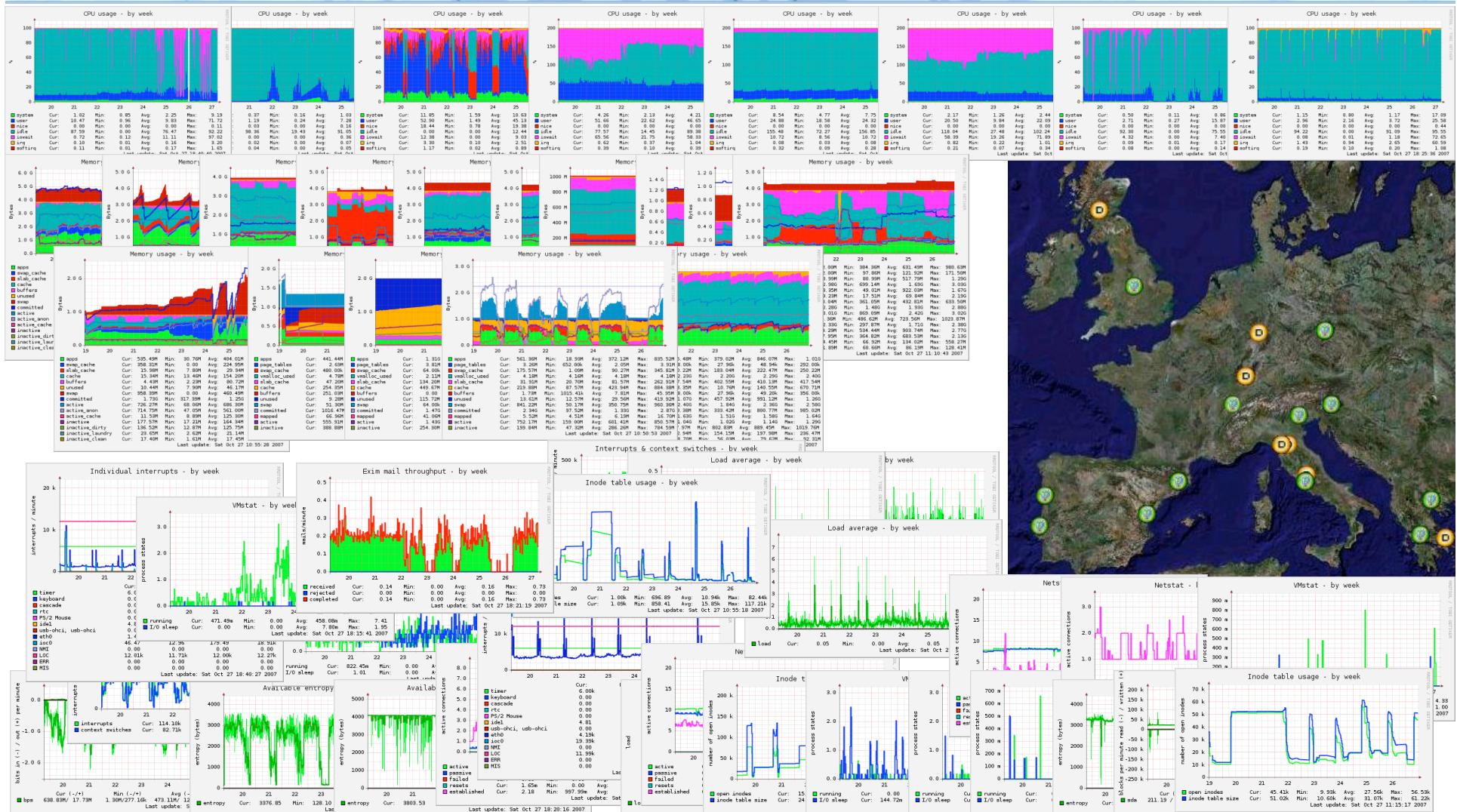


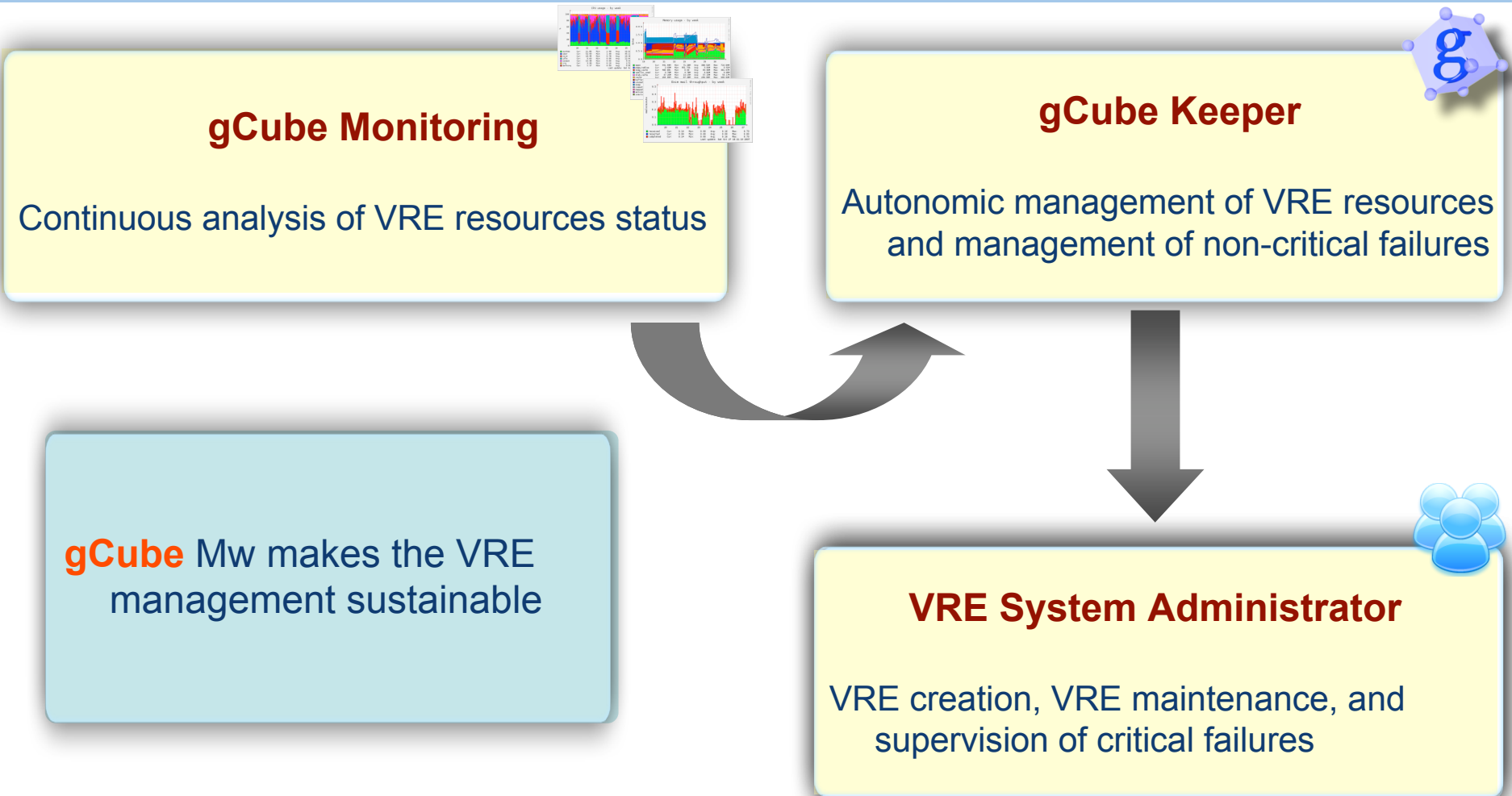
Diligent

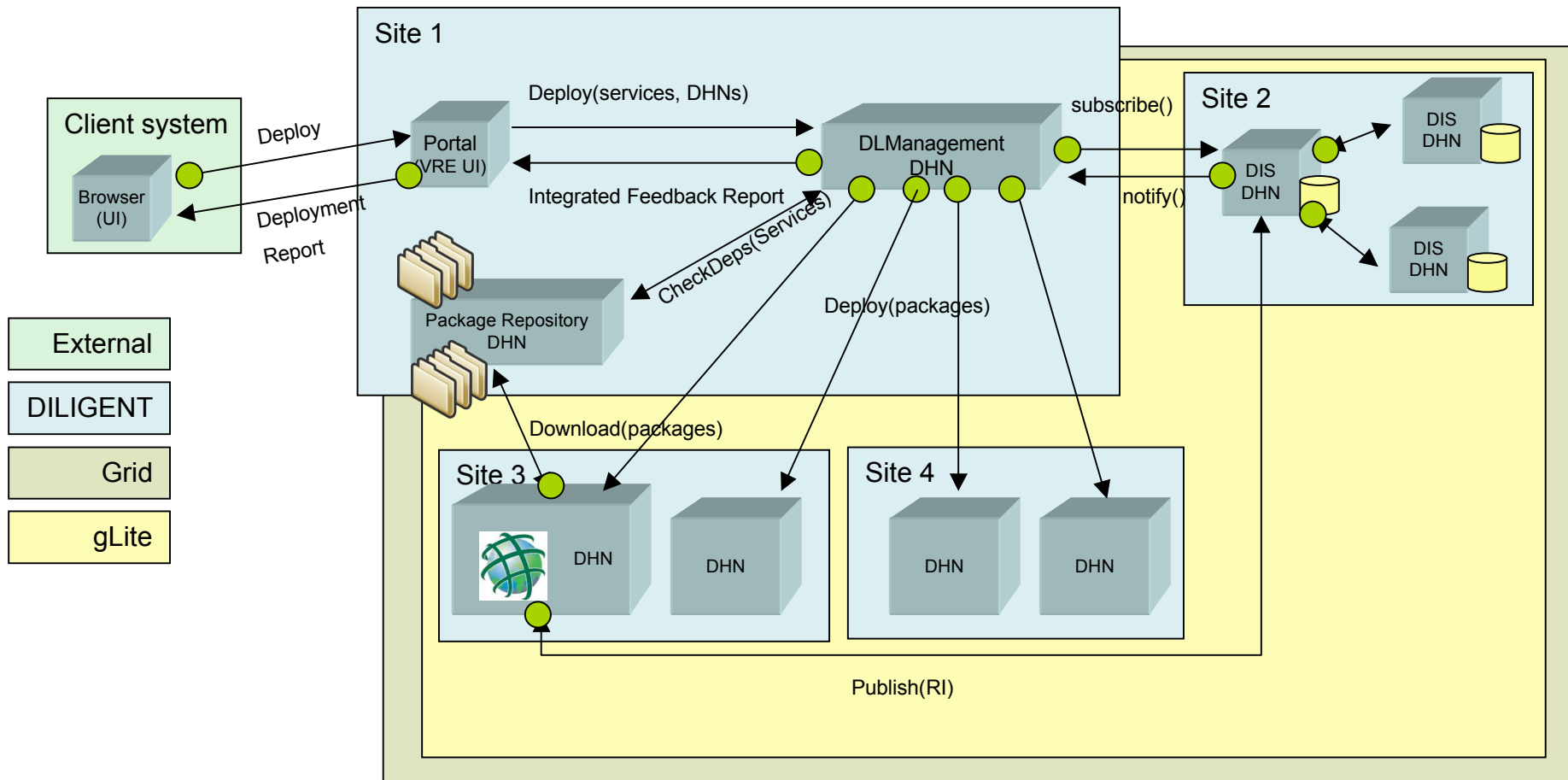
From Digital Objects  
to Content across  
eInfrastructures

Pasquale Pagano  
CNR-ISTI









# ImpECT VRE



Diligent

From Digital Objects  
to Content across  
eInfrastructures



- 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

**GeoNetwork**  
[Find and analyze geo-spatial data]

Search for Data and Information

Recent Additions 555

- Global bovine density (2005)
- Cattle density for sub-Saharan Africa (2005)
- Global cattle density (2005)
- Coastline of the World (Vmap0)
- Non Perennial Water Courses (2005)
- Perennial Water Courses (2005)
- Non Perennial Inland Water Courses (2005)

www.fao.org/geonetwork

~4700 global data set available

**Medspiration.org**  
The European Sea Surface Temperature Project

The Medspiration Project - An Introduction

Sea Surface Temperature

Satellite Measurements

www.medspiration.org

daily data sets available

**GEMES**  
Global Monitoring for Environment and Security

Overview

Towards Services

Management

Achievements

National Activities

Library

Newsletter

www.gmes.int

Reference doc Metadata, services

**CEOS**  
Committee On Earth Observation Satellites

International Directory Network (IDN)

Welcome to the CEOS International Directory Network - A Gateway to a World of Earth Science Data

idn.ceos.org

ES Thesaurus ~30000 objects

**eoGRID.esrin.esa.int**

Services page of European Space Agency (ESA) Earth Observation (EO) products and services, data and computer elements that are currently available according to level and privileges (available in the General Information tab). For any request or information requests please feel free to email us.

General Information

eoGRID.esrin.esa.int

environmental data / reports

**eea.eu.int**

EU must take immediate action on Kyoto target

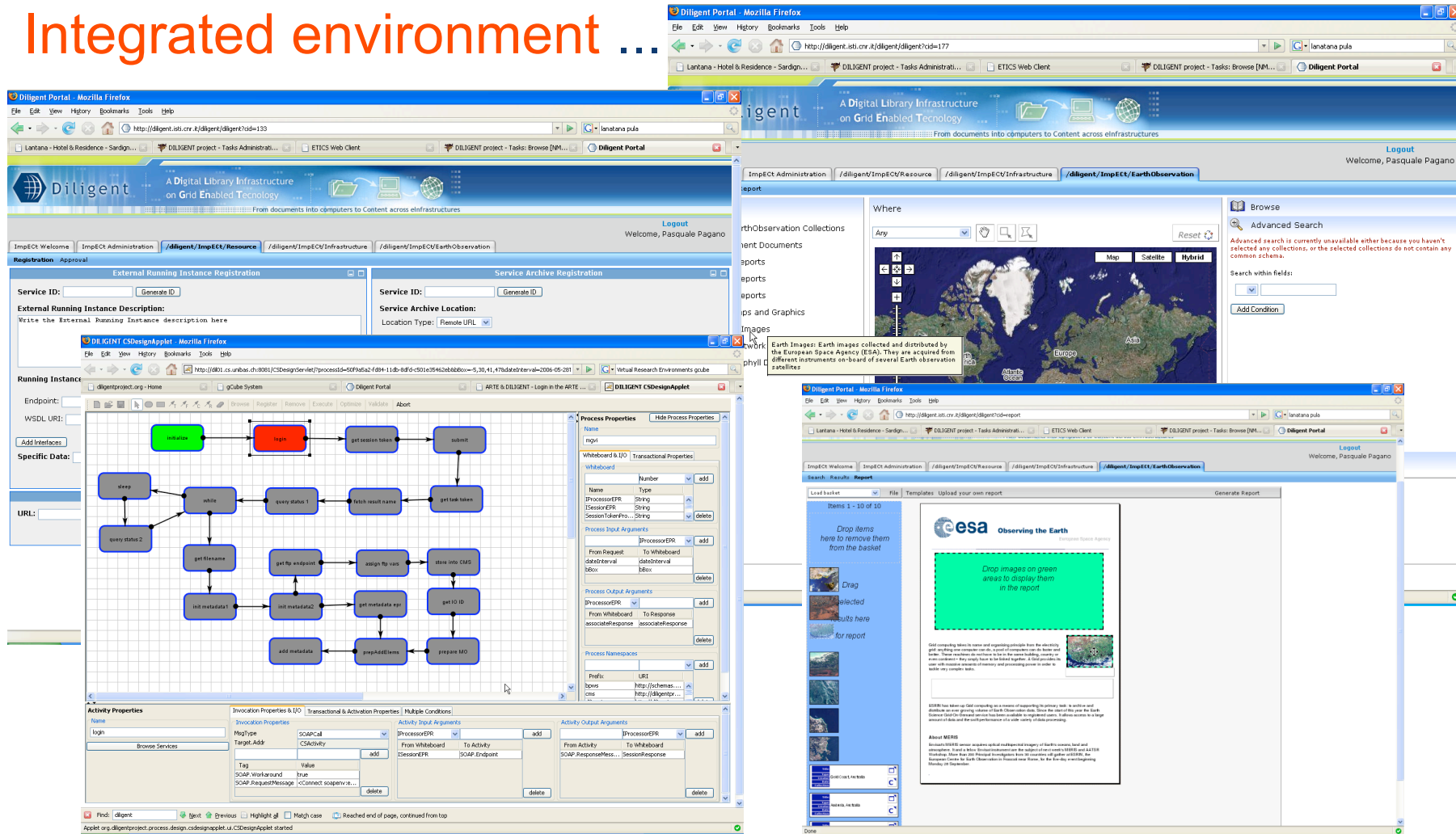
27 Oct 2006

Greenhouse gas emission trends and projections in Europe 2006

Annual Climate Change Report

www.eea.eu.int

## Integrated environment



The screenshot displays an integrated environment with several browser windows and a workflow diagram. The main window shows the Diligent portal with a navigation menu and a search interface. A secondary window displays a workflow diagram with various activity nodes and properties. A third window shows a search results page for Earth Observation data from ESA.

**Workflow Diagram:**

```

    graph TD
      init[init] --> login[login]
      login --> getSession[getSession]
      getSession --> submit[submit]
      submit --> getTask[getTask]
      getTask --> queryStatus1[query status 1]
      queryStatus1 --> fetchResult[fetch result]
      fetchResult --> getTaskToken[get task token]
      getTaskToken --> queryStatus2[query status 2]
      queryStatus2 --> write[write]
      write --> getFileName[get filename]
      getFileName --> getEndpoint[get endpoint]
      getEndpoint --> assignToUser[assign to user]
      assignToUser --> showInfo[show info]
      showInfo --> initMetadata1[init metadata 1]
      initMetadata1 --> initMetadata2[init metadata 2]
      initMetadata2 --> getMetadata[get metadata]
      getMetadata --> getID[get ID]
      getID --> addMetadata[add metadata]
      addMetadata --> prepAddItems[prep add items]
      prepAddItems --> prepareBIO[prepare BIO]
  
```

**Search Results Page:**

Search Results: Report

Items 1 - 10 of 10

Drop items here to remove them from the basket

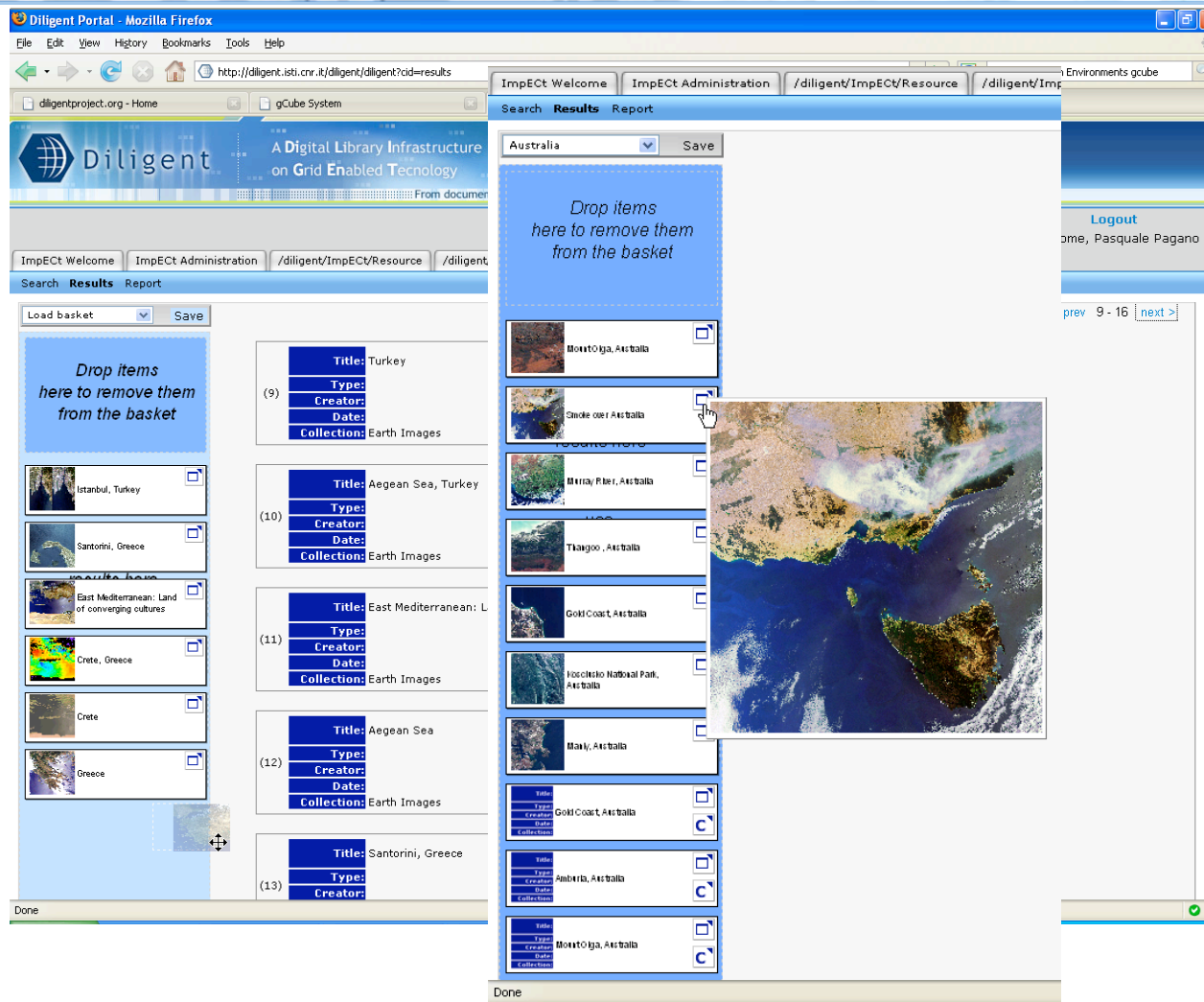
Drop images on green areas to display them in the report

esa Observing the Earth

ESA has taken up GDI complying to a process of responding to industry, earth to earth and distribute an ever growing volume of Earth Observation data. Since the start of this year the Earth Observation GDI for Global users has been available as integrated system. System access to a large amount of data and the best performance of a wide variety of data processing.

... 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,
- ...



The screenshot shows the ImpEct private workspace interface. The browser window is titled "Diligent Portal - Mozilla Firefox" and displays the URL "http://diligent.isti.cnr.it/diligent/diligent?cid=results". The interface includes a search bar with "Australia" selected, a "Load basket" section on the left, and a central list of search results. The results list includes items such as "Turkey", "Aegean Sea, Turkey", "East Mediterranean: Land of converging cultures", "Aegean Sea", and "Santorini, Greece". A right-hand sidebar contains a "Drop items here to remove them from the basket" area with a list of items including "Montolga, Australia", "Smoke over Australia", "Murray River, Australia", "Tiaajoo, Australia", "Gold Coast, Australia", "Fossiliferous National Park, Australia", "Manki, Australia", "Gold Coast, Australia", "Amberley, Australia", and "Montolga, Australia". A large satellite image of Australia is displayed in the background of the right sidebar.

- 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

- **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

- **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

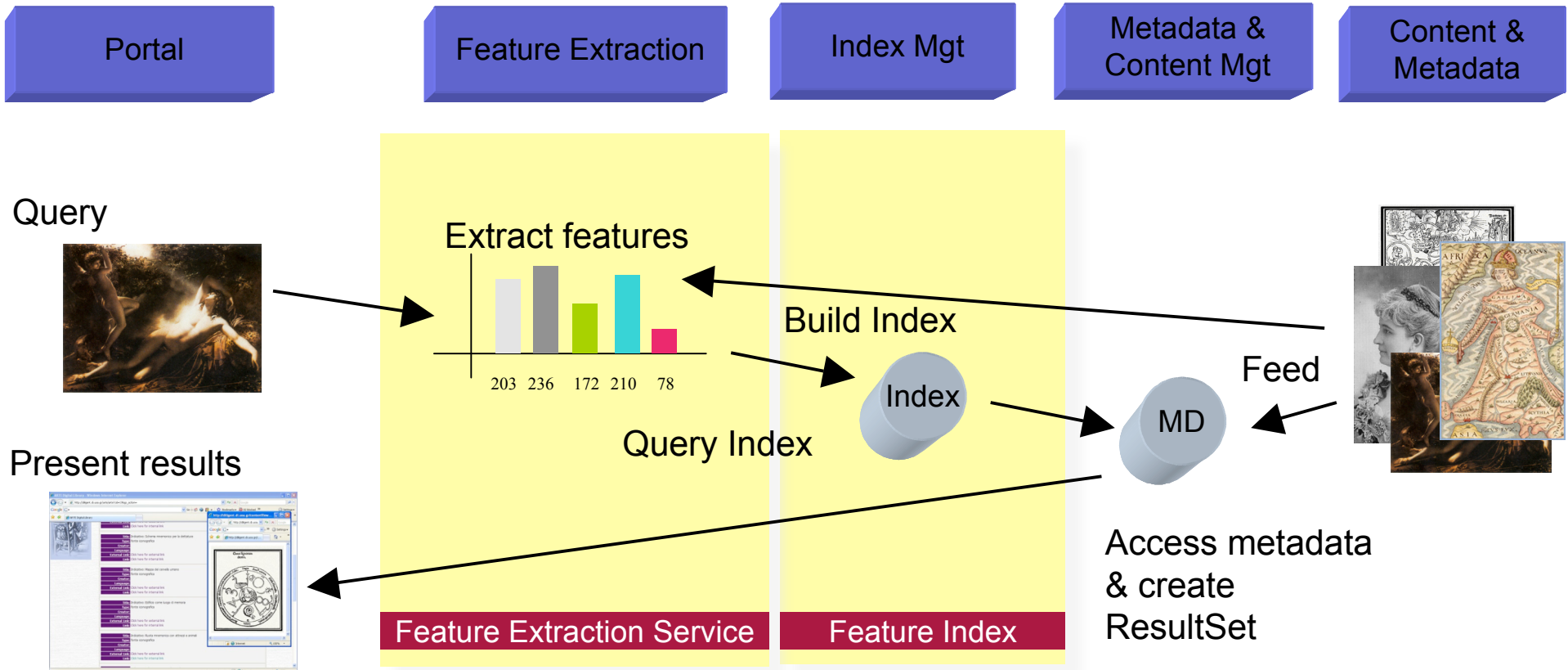
Retrieval of Distributed **Information**

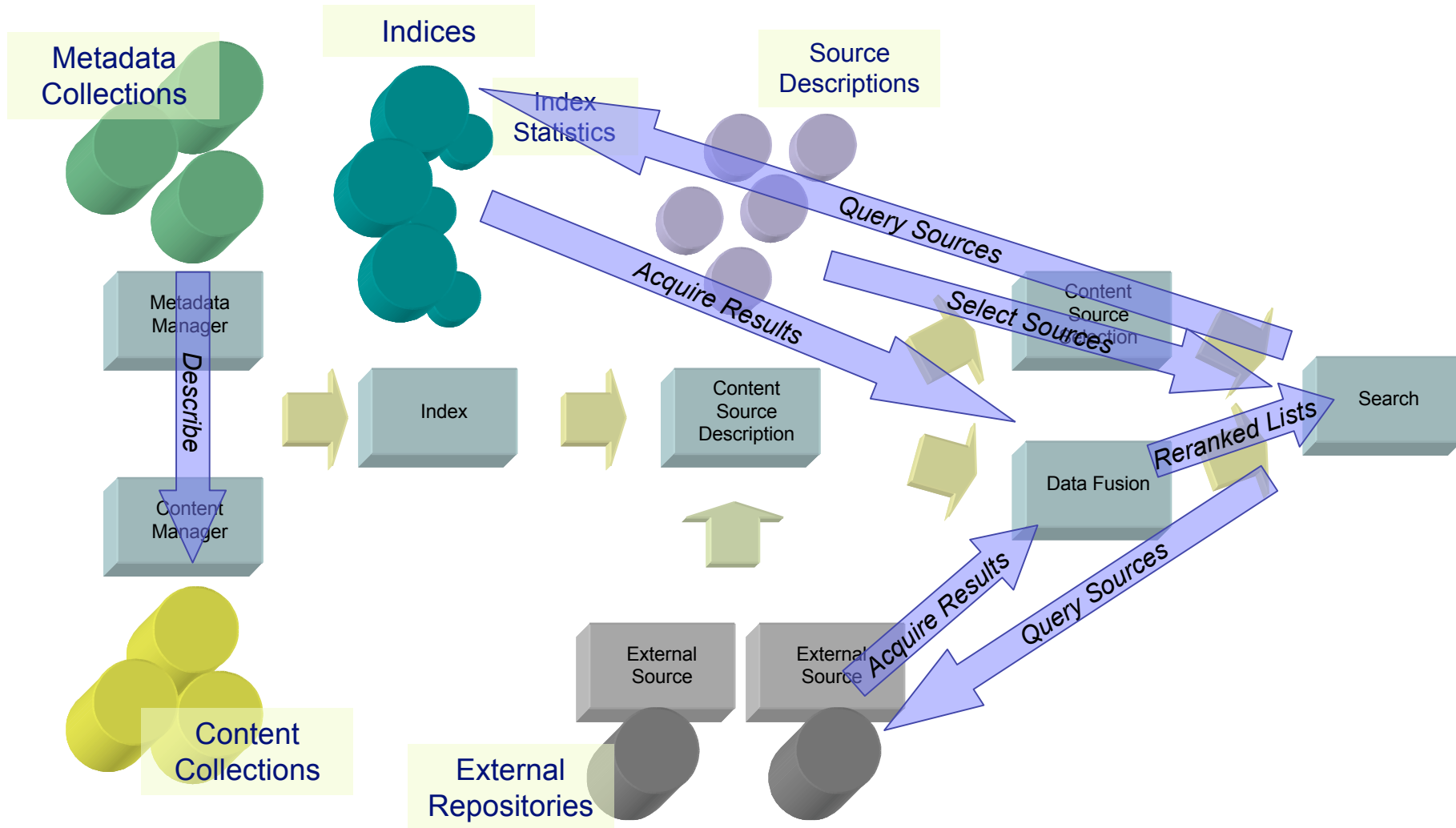
Distributed **Retrieval** of Information

- **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



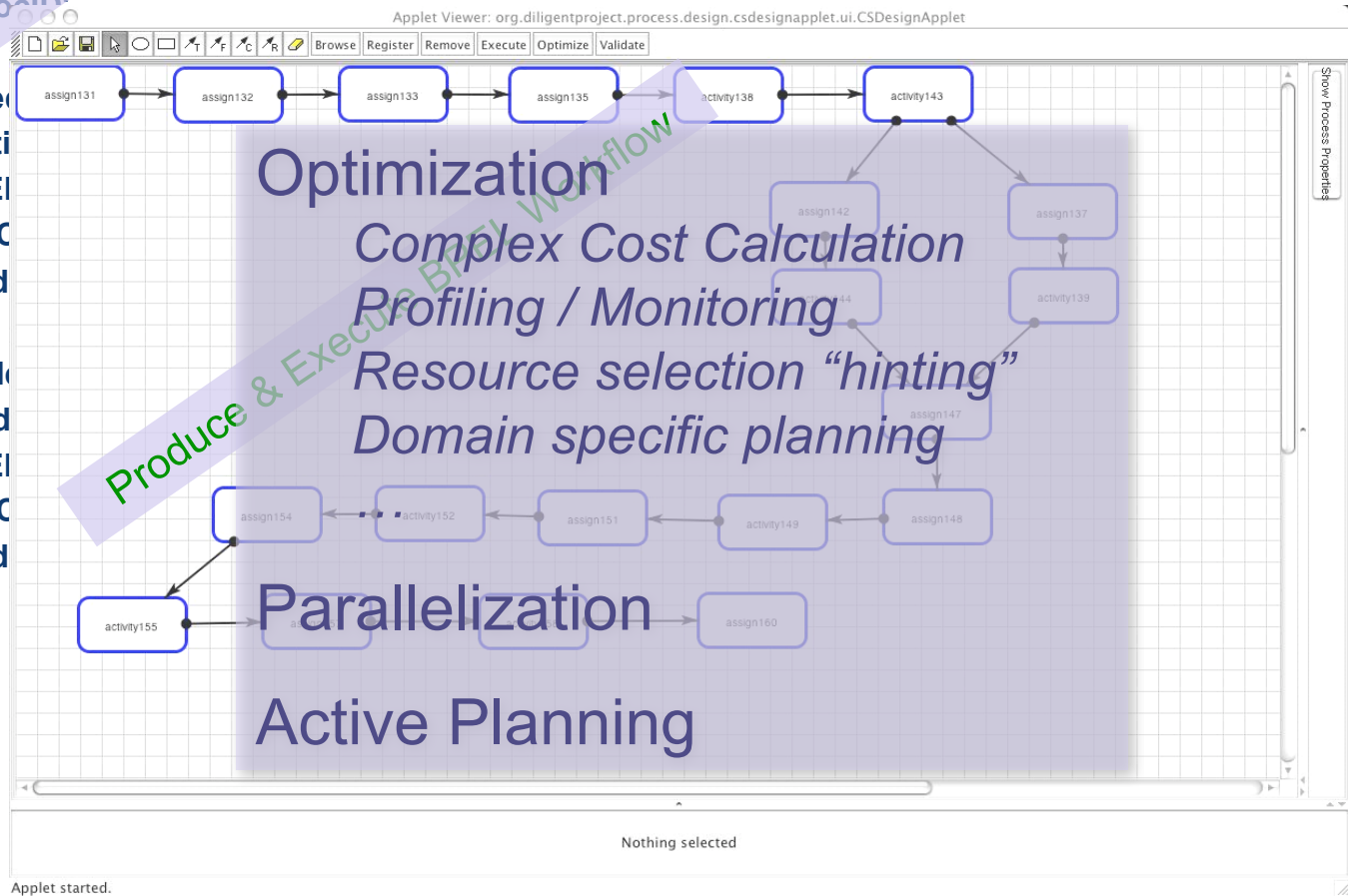


- Numerous Search services, for info retrieval & processing
  - Structured data and XML processing (scanners, sorters, joiners, filterers, transformers, retrievers)
  - Lookups (indices, FT indices, XML indices, Geo indices)
  - Content-based searches
  - External source probes
  - Fusion / Merging of 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

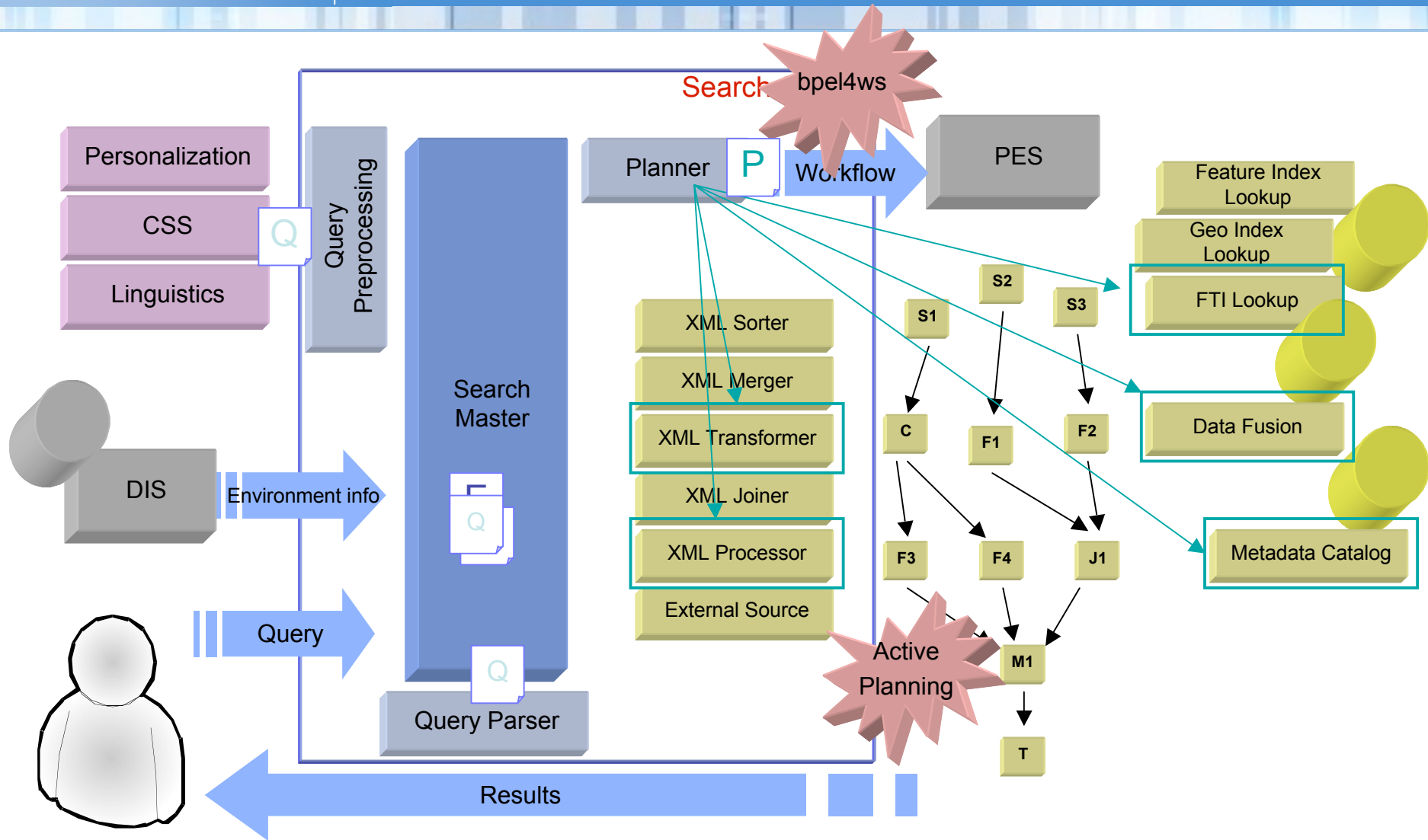
Query

and (field  
 by 'd  
 in 'E  
 on 'C  
 as 'd



```
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')
              )
            )
          )
        )
      )
    )
  )
```

- **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





Diligent

From Digital Objects  
to Content across  
eInfrastructures



**from theory ...  
... to reality**