

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



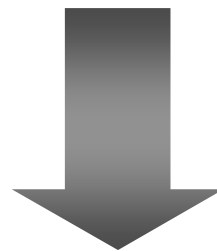
Diligent

From Digital Objects
to Content across
eInfrastructures

Donatella Castelli
ISTI-CNR



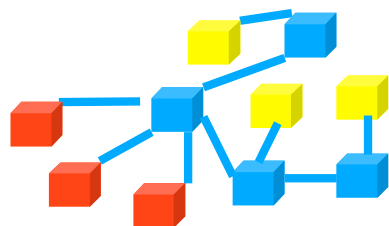
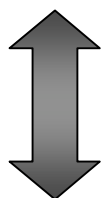
- DLs are evolving into “Virtual Research Environments” (Collaboratoria)
 - Distributed frameworks for carrying out **cooperative activities** like “in silico experiments”, data analysis and processing, production of new knowledge using specialised tools
 - Largely based on **retrieval and access** of always updated knowledge from diverse heterogeneous content sources
 - The knowledge produced is **preserved** and **made available** for other usages inside and outside the VRE



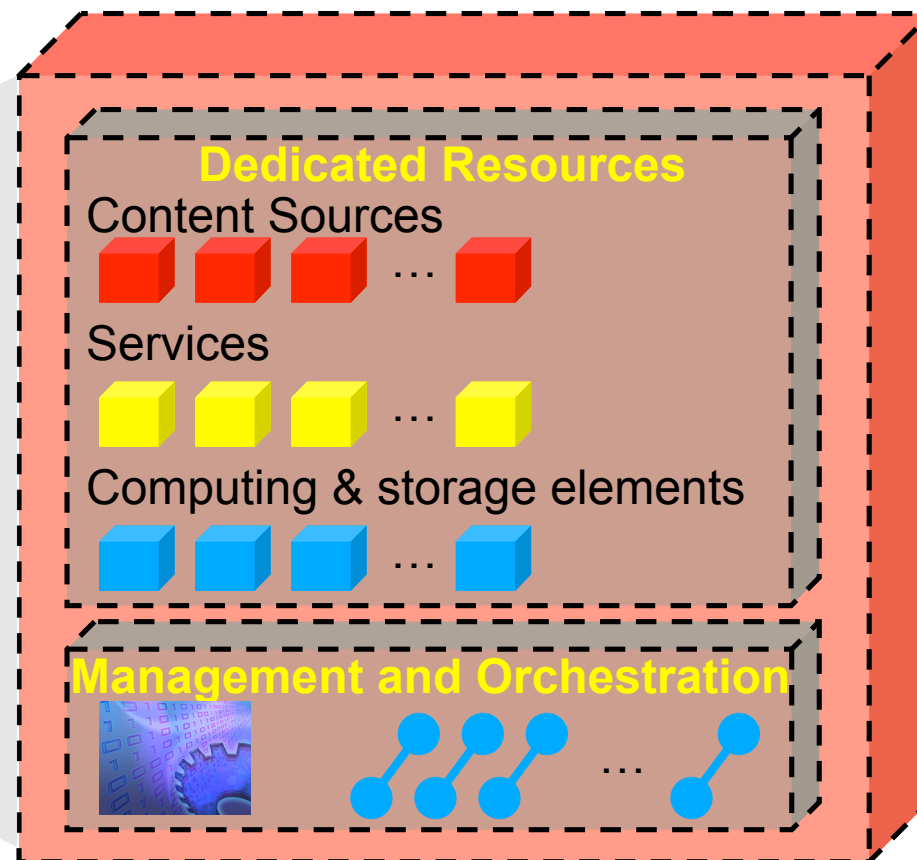
Highly dynamic, created and dismissed on-demand

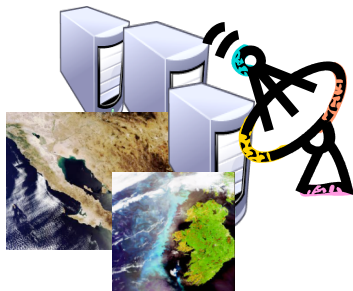


Based on specialised tools which support the generation of new knowledge

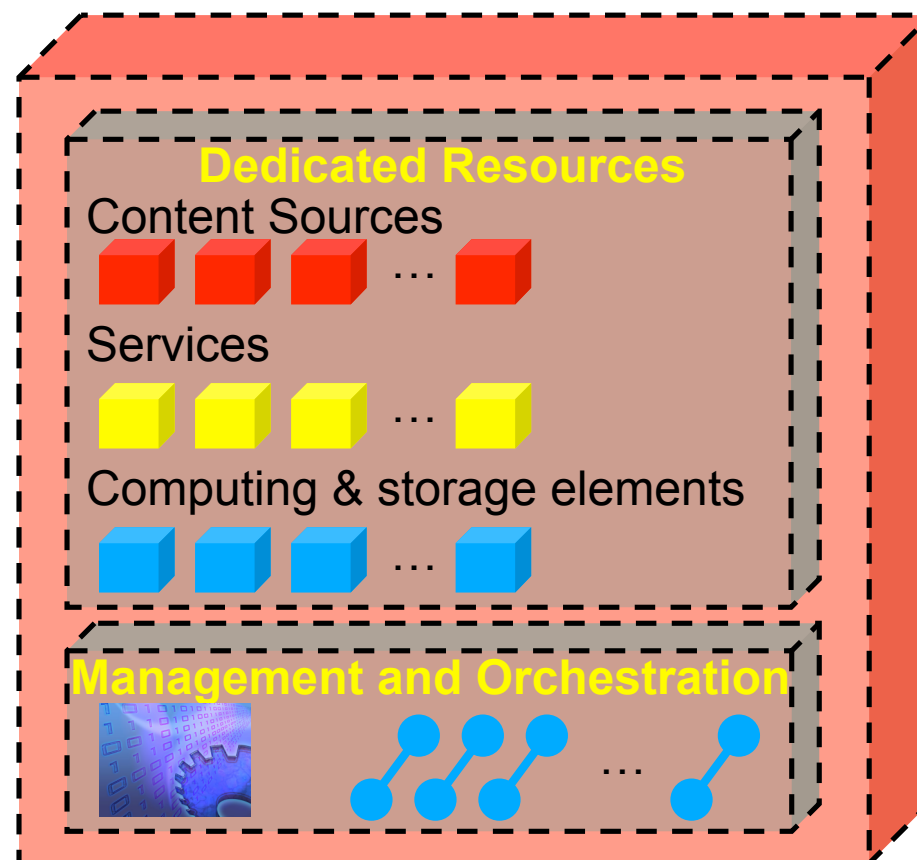


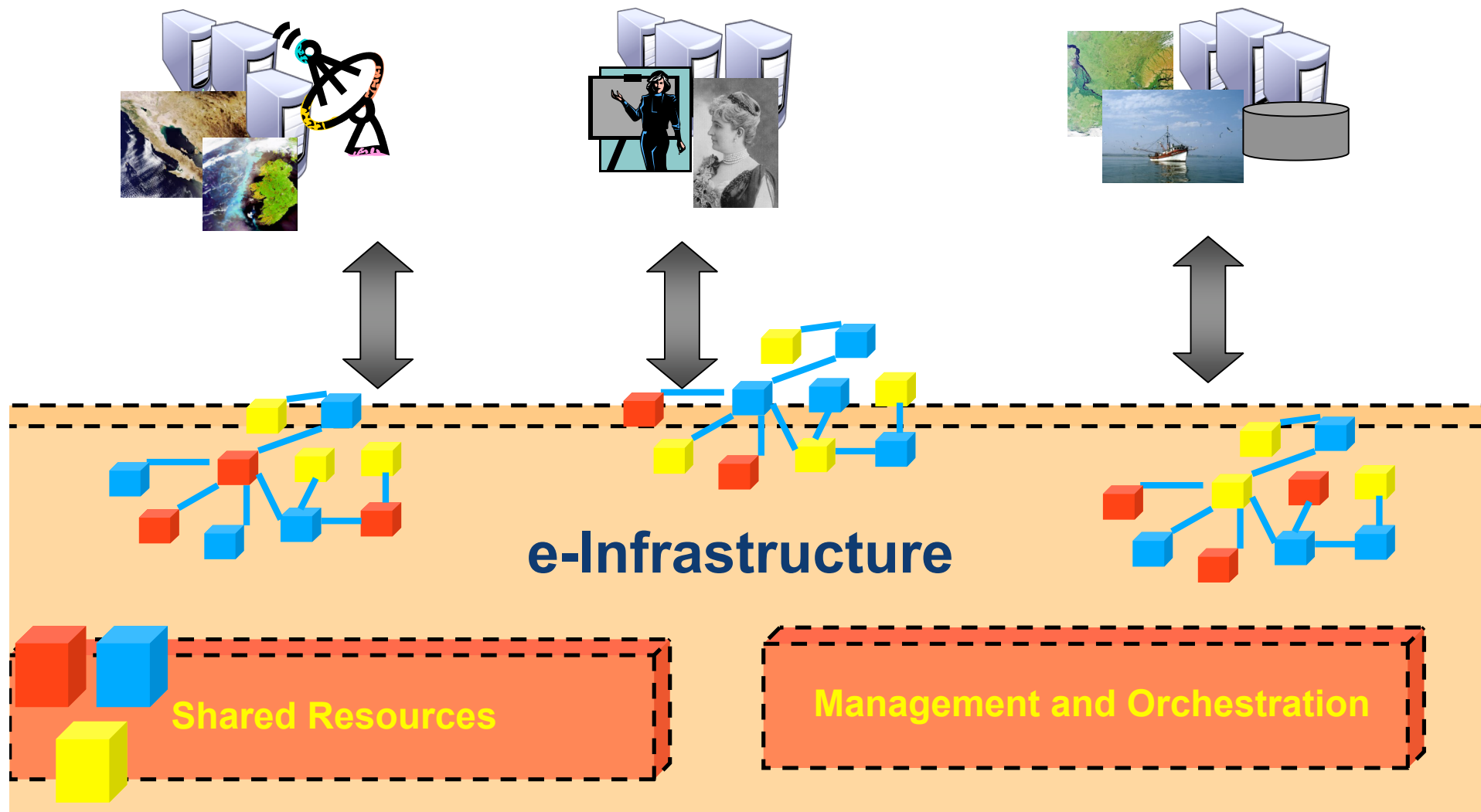
VRE System





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



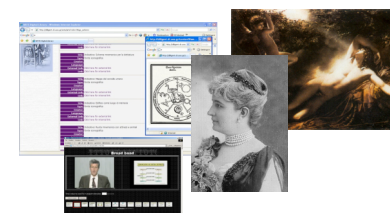


- **Infrastructure sustainability**
 - Mechanisms for reducing the cost of the infrastructure mng
- **Supported VREs**
 - Flexible and high quality solutions for satisfying the needs of many different applications domains
 - Simple procedures for creating VREs

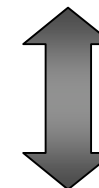
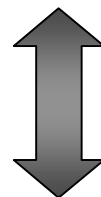
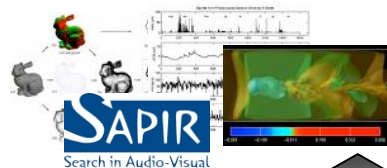
ImpECt Environmental Monitoring



ARTE Education in the Humanities



SAPIR-enabled AV search

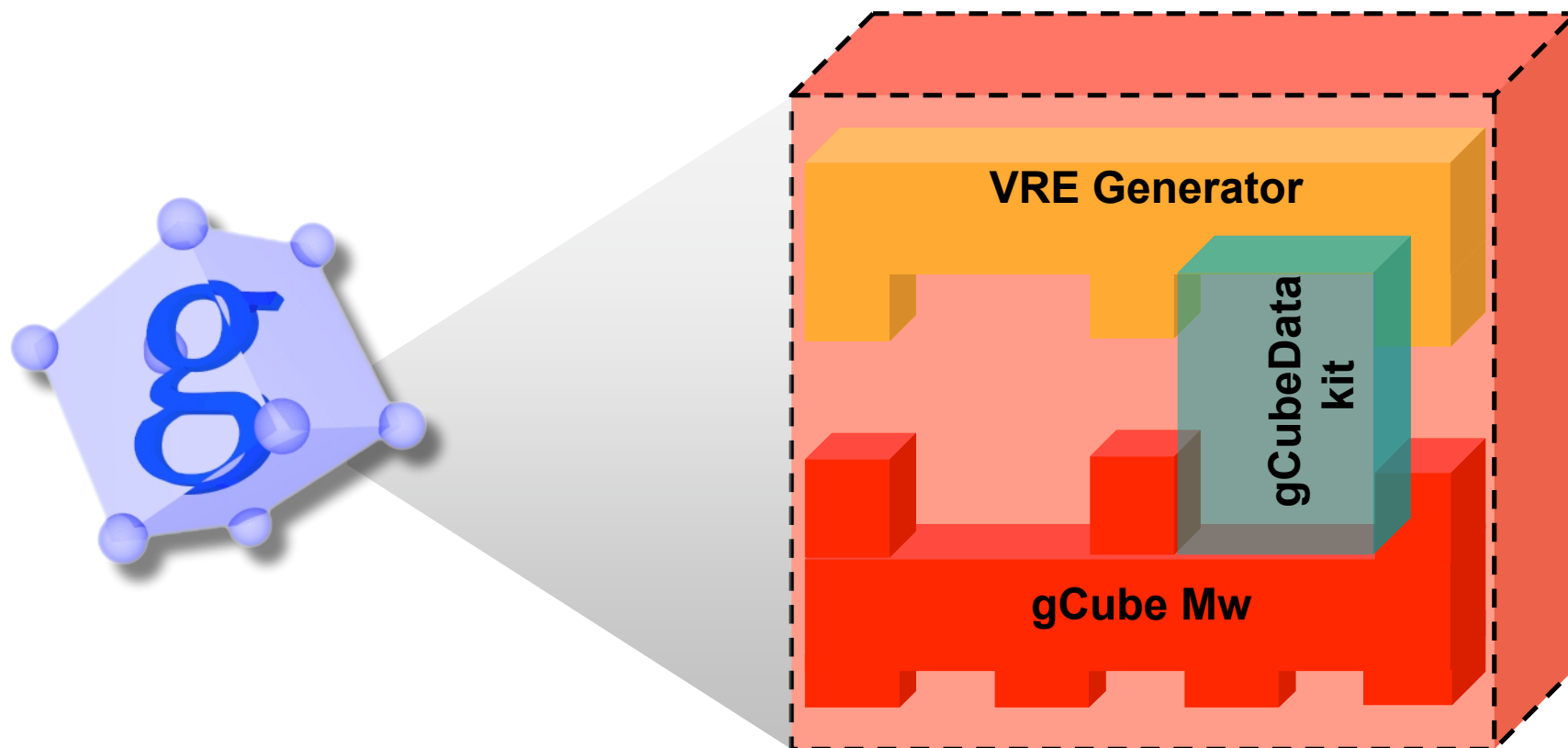


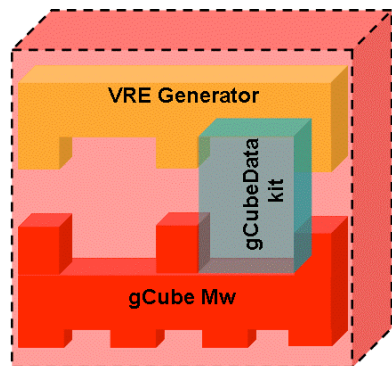
DILIGENT
Infrastructure



gCube System







Simplifies the infrastructure management

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



Resource

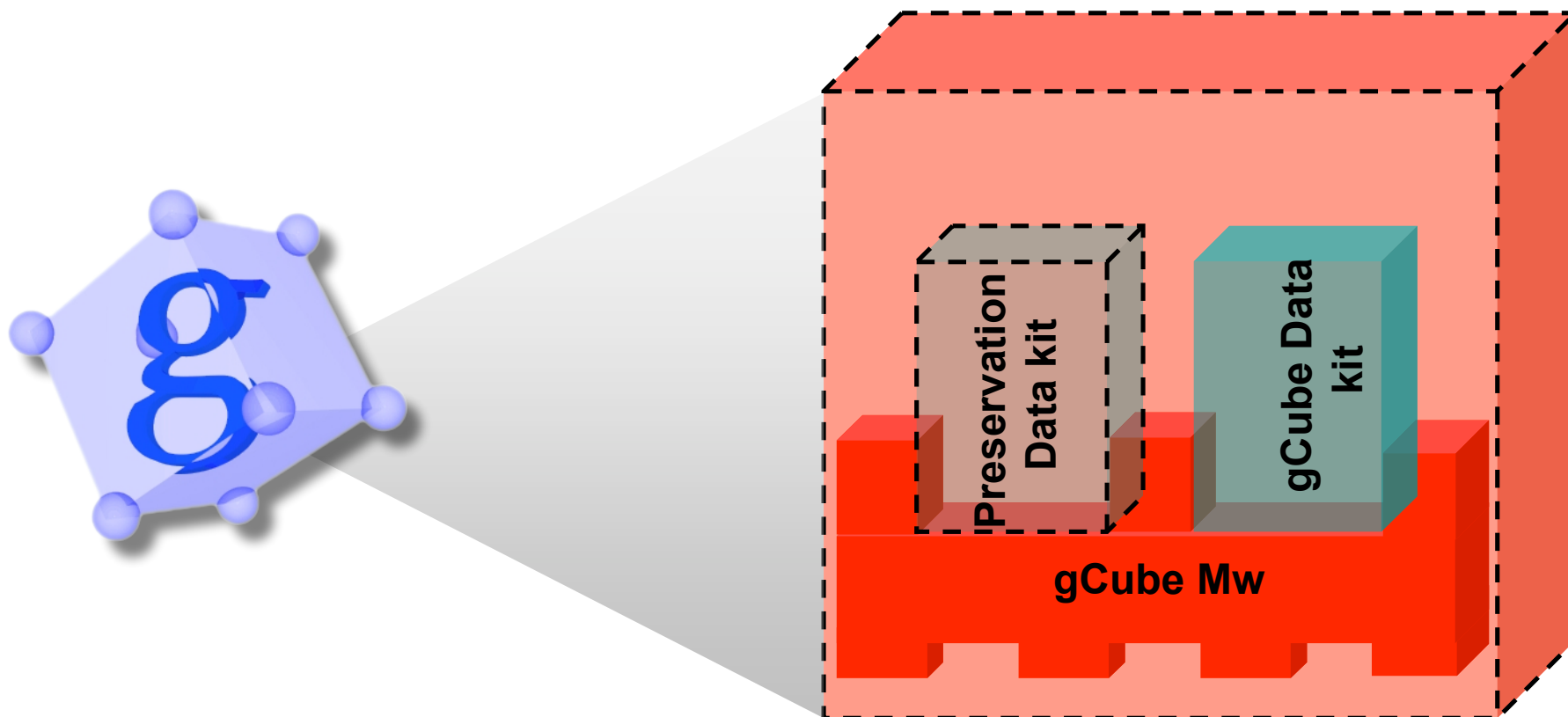


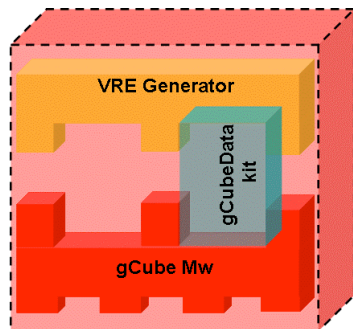
Service

Content Source



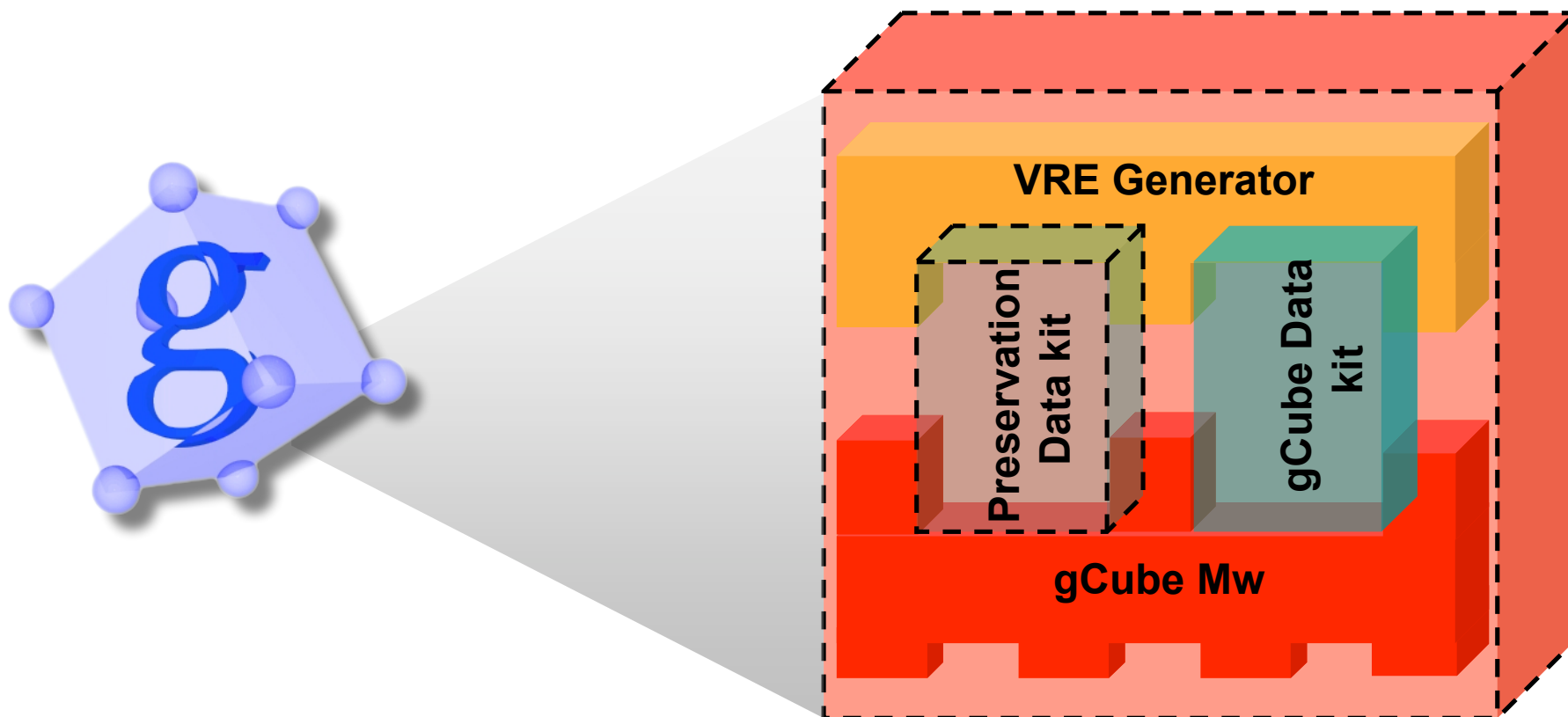
Comp&Storage

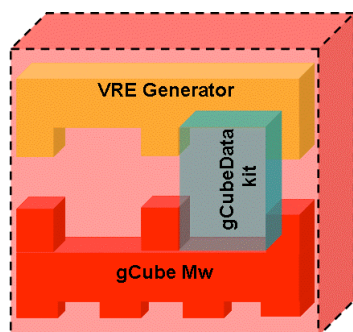




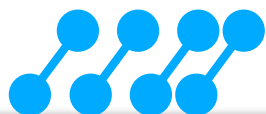
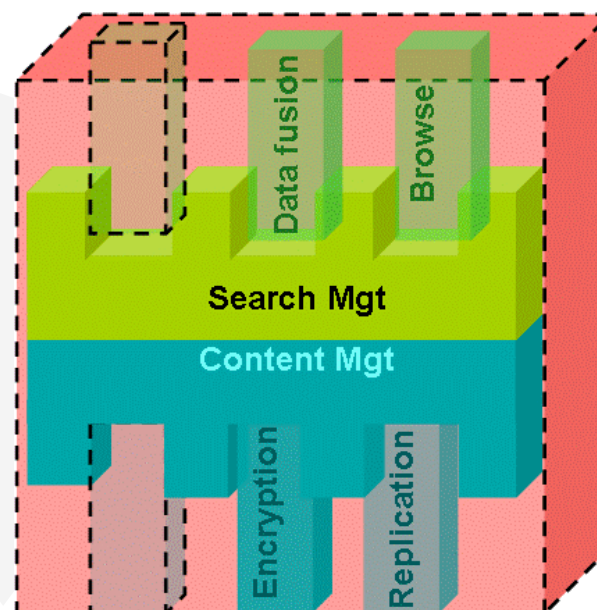
Simplifies the construction of a VRE system

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





**Provides flexible search
and management functionality**

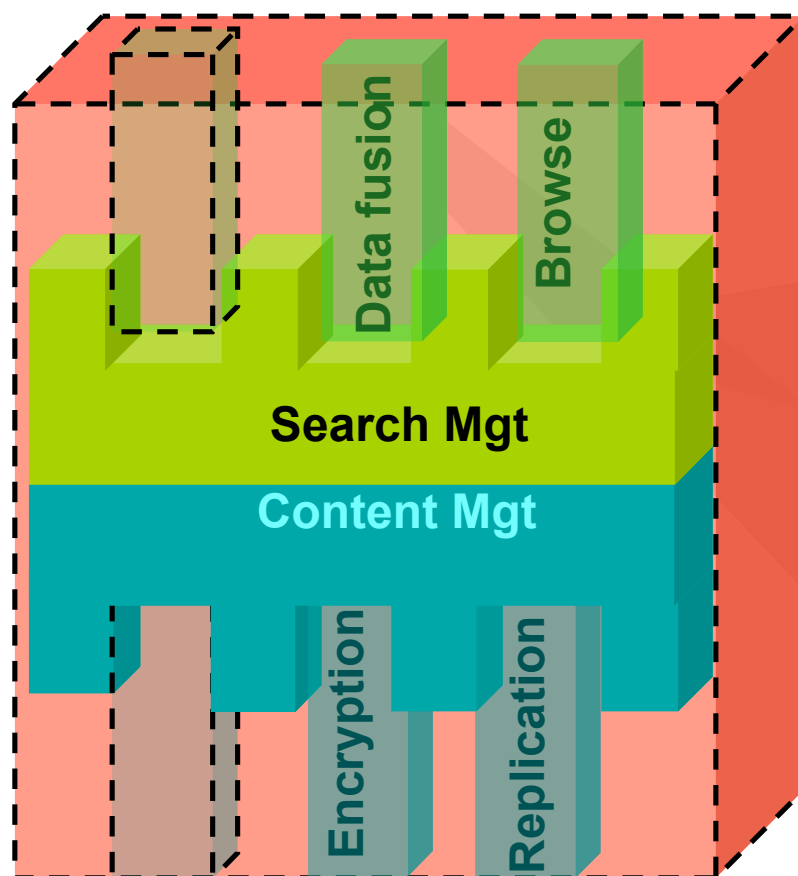


Data Fusion

Browse

Source sel.

Feature extr.



Most important framework for
Information Spaces

Most important functionality /
service in Information Access

VRE Management

Pasquale Pagano
CNR-ISTI

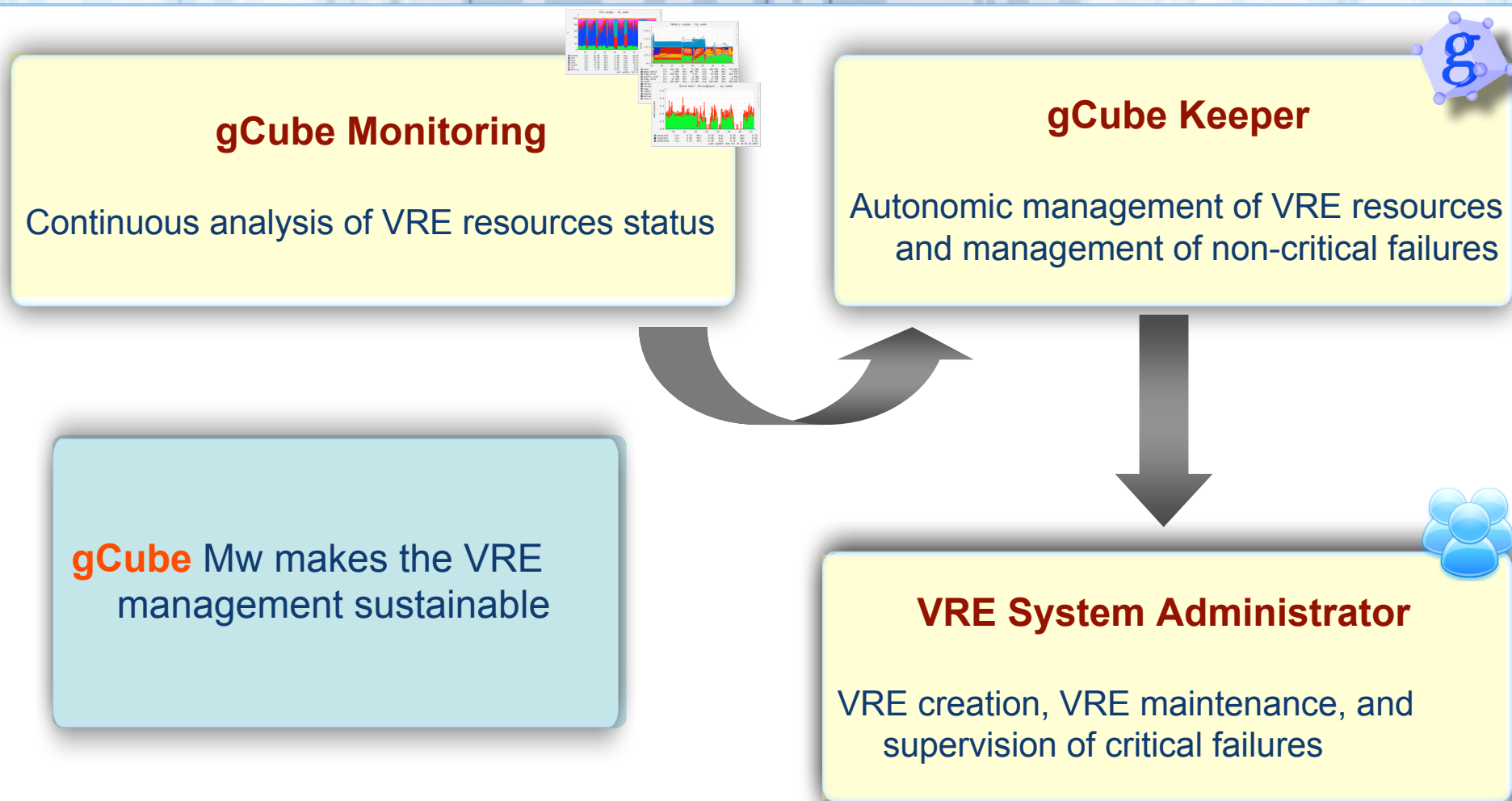


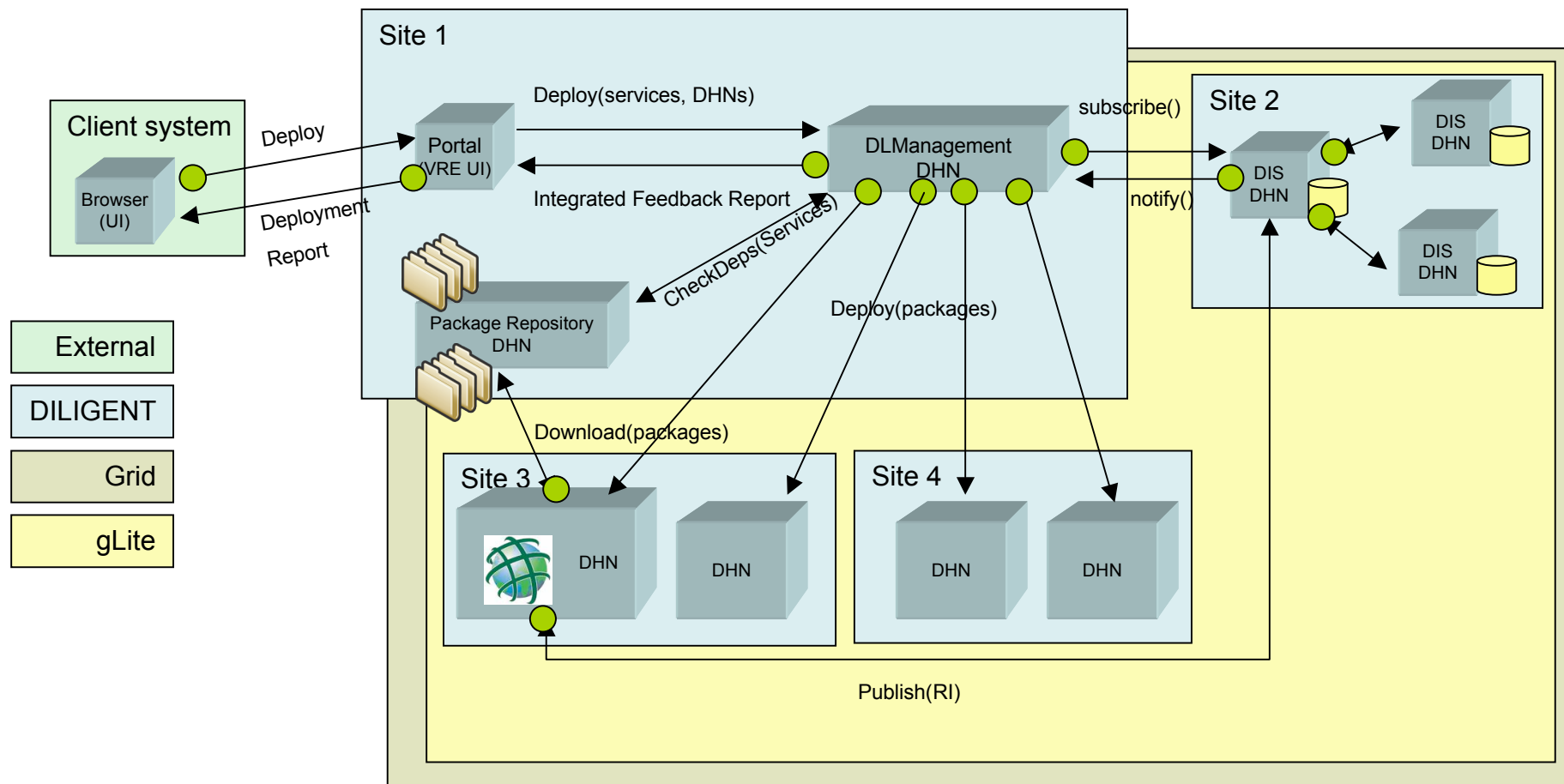
Diligent

From Digital Objects
to Content across
eInfrastructures









ImpECt VRE



Diligent

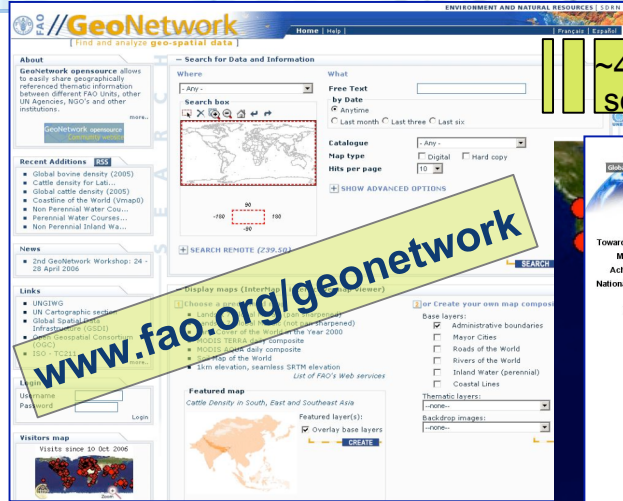
From Digital Objects
to Content across
eInfrastructures



- 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



Data Sources from the Web...

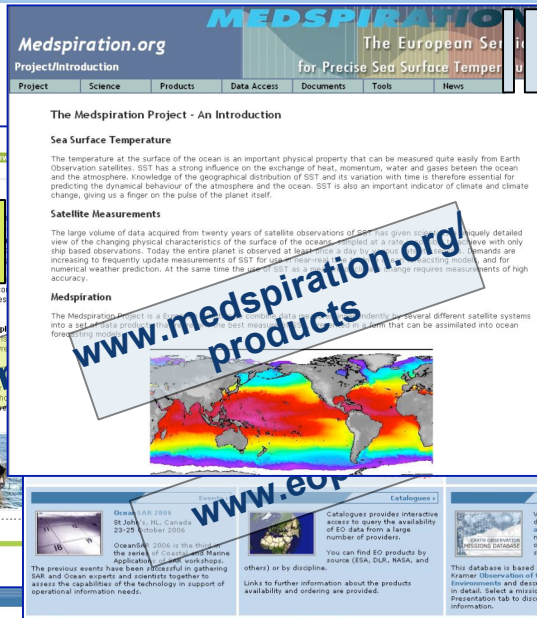


~4700 global data set available

www.fao.org/geonetwork



www.gmes.int

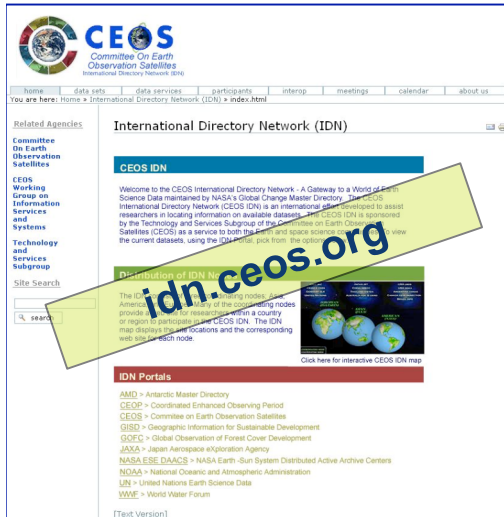


daily data sets available

Reference doc Metadata, services

www.medspiration.org

www.eea.eu.int



ES Thesaurus ~30000 objects

idn.ceos.org



eogrid.esrin.esa.int



environmental data / reports

www.eea.eu.int



Diligent for ImpEct

Integrated environment

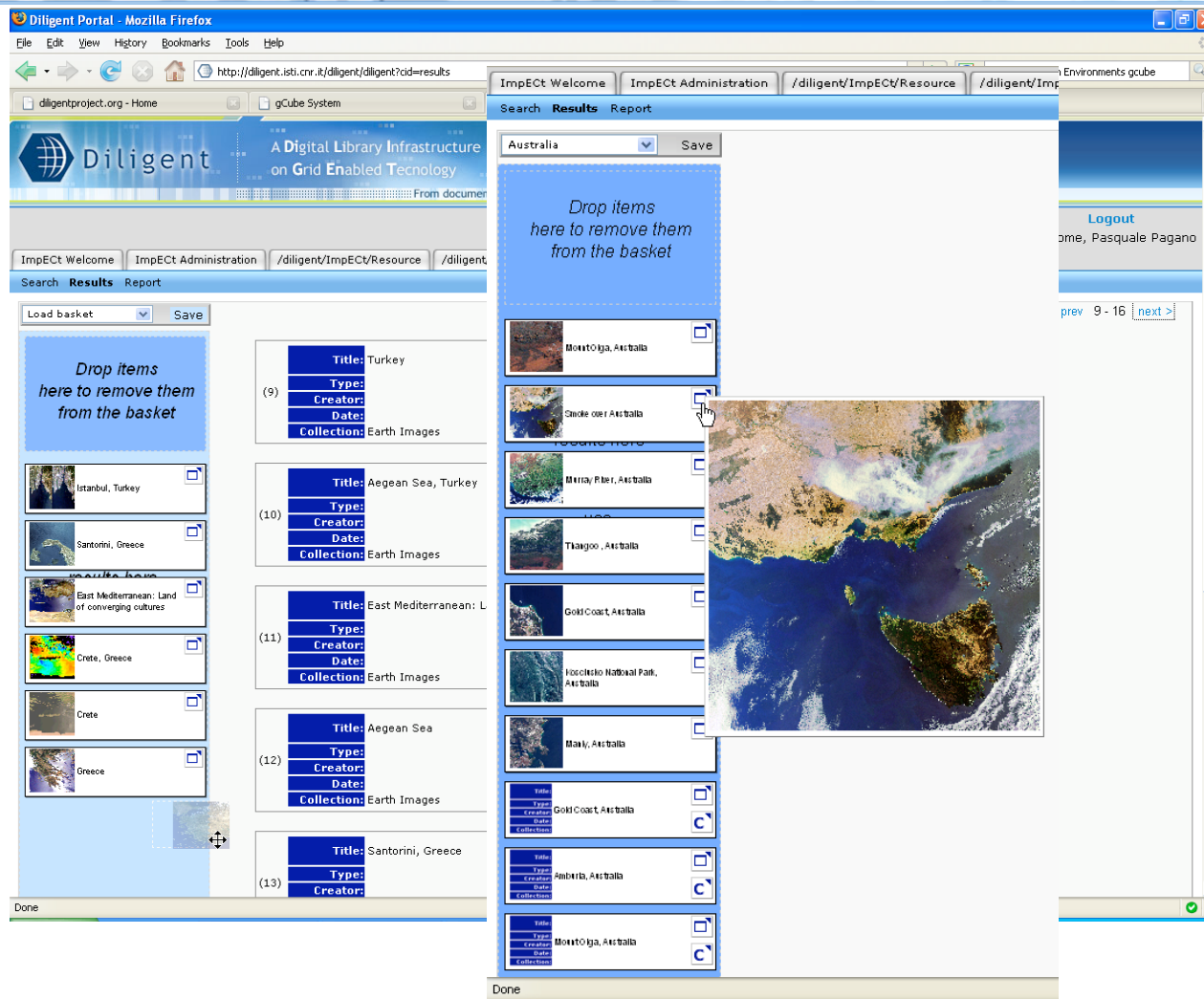
The screenshot displays the Diligent for ImpEct integrated environment, which is a web-based infrastructure for managing and processing data. The interface is divided into several sections:

- Top Navigation Bar:** Includes the Diligent logo, a search bar, and links to various services and resources.
- Left Sidebar:** Contains navigation links for ImpEct Administration, ImpEct Resource, ImpEct Infrastructure, and ImpEct Earth Observation.
- Main Content Area:** Displays a workflow diagram with various activities and data flows. The workflow includes steps like 'login', 'get session token', 'submit', 'query status 1', 'fetch result name', 'get task token', 'query status 2', 'get filename', 'get file endpoint', 'assign file size', 'store into CMS', 'init metadata1', 'init metadata2', 'get metadata expr', 'get ID ID', 'add metadata', 'prepareAddItems', and 'prepare MD'. The diagram is organized into a grid with arrows indicating the flow of data and control.
- Right Sidebar:** Contains a 'Where' section with a map of Europe and a 'Browse' section with a search bar and filters.
- Bottom Section:** Displays a list of items (1-10 of 30) and a detailed view of a specific item, including a description and a list of related items.

The interface is designed to be user-friendly and efficient, allowing users to manage their data and workflows effectively. The workflow diagram is a key feature, providing a visual representation of the data processing pipeline. The map and search functionality are also important for locating and retrieving data. The bottom section provides a detailed view of the data, allowing users to explore and analyze it in depth.

... 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 in Mozilla Firefox. The browser window displays the Diligent portal with a search results page. The interface includes a 'Load basket' button and a 'Drop items here to remove them from the basket' instruction. A list of search results is displayed, including items like 'Istanbul, Turkey', 'Santorini, Greece', 'East Mediterranean: Land of converging cultures', 'Crete, Greece', 'Crete', 'Greece', 'Aegean Sea, Turkey', 'Aegean Sea', 'Santorini, Greece', 'Australia', 'Cinder cone / Australia', 'Murray River, Australia', 'Tiajoo, Australia', 'Gold Coast, Australia', 'Pioscedo National Park, Australia', 'Manki, Australia', 'Gold Coast, Australia', 'Amberia, Australia', and 'Montolga, Australia'. A large satellite image of the Aegean Sea region is shown on the right.

- 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

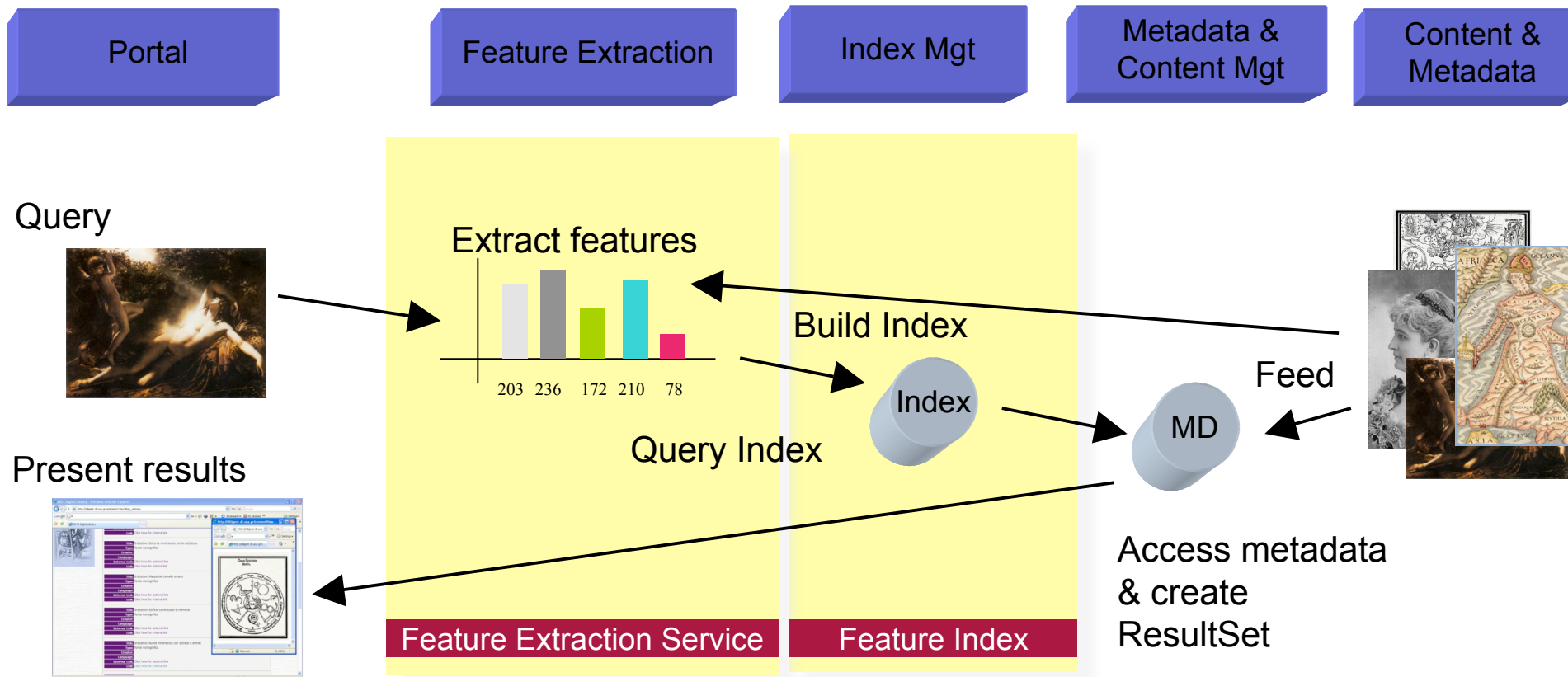
Retrieval of Distributed **Information**

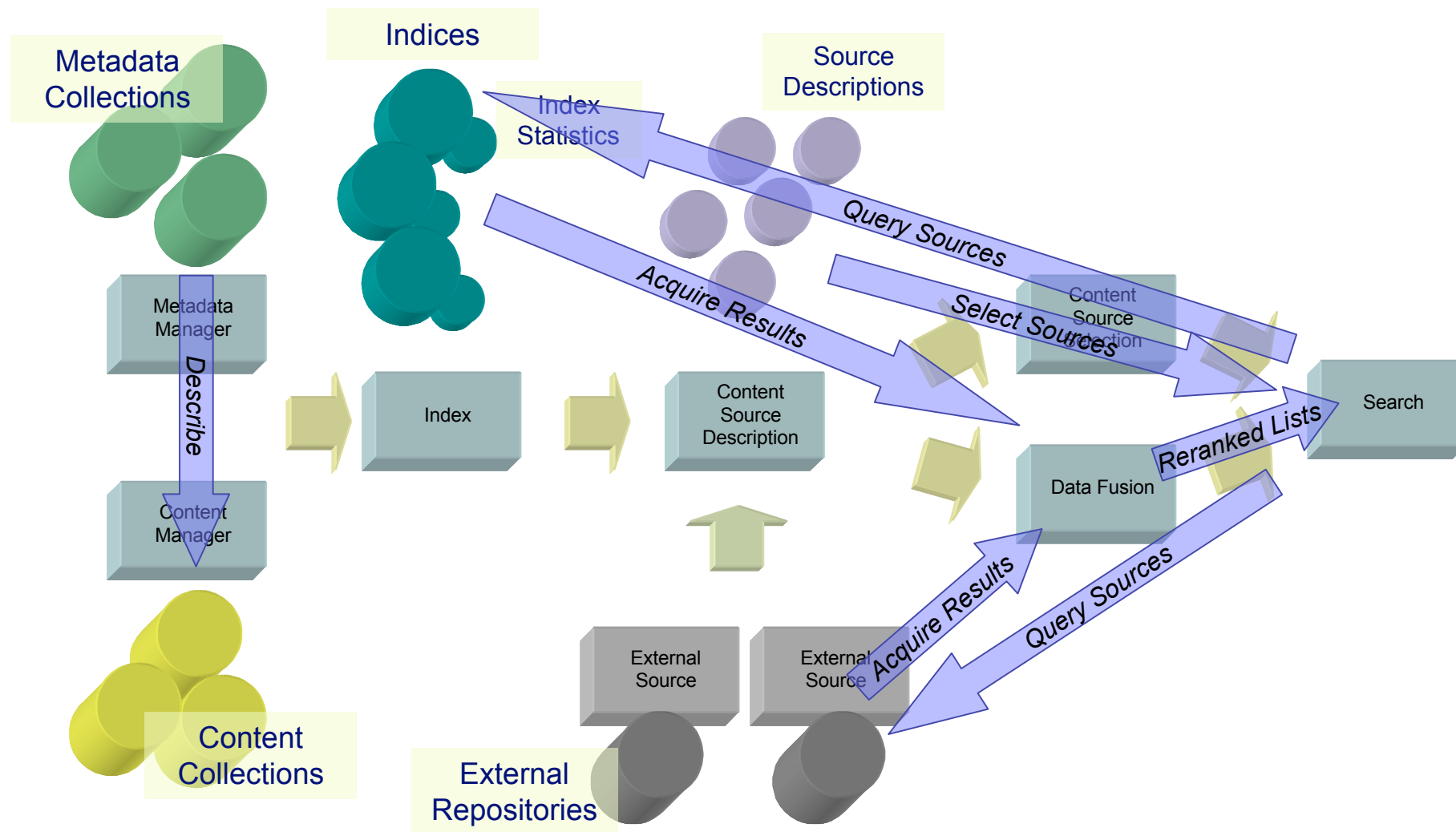
Distributed **Retrieval** of Information

- **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



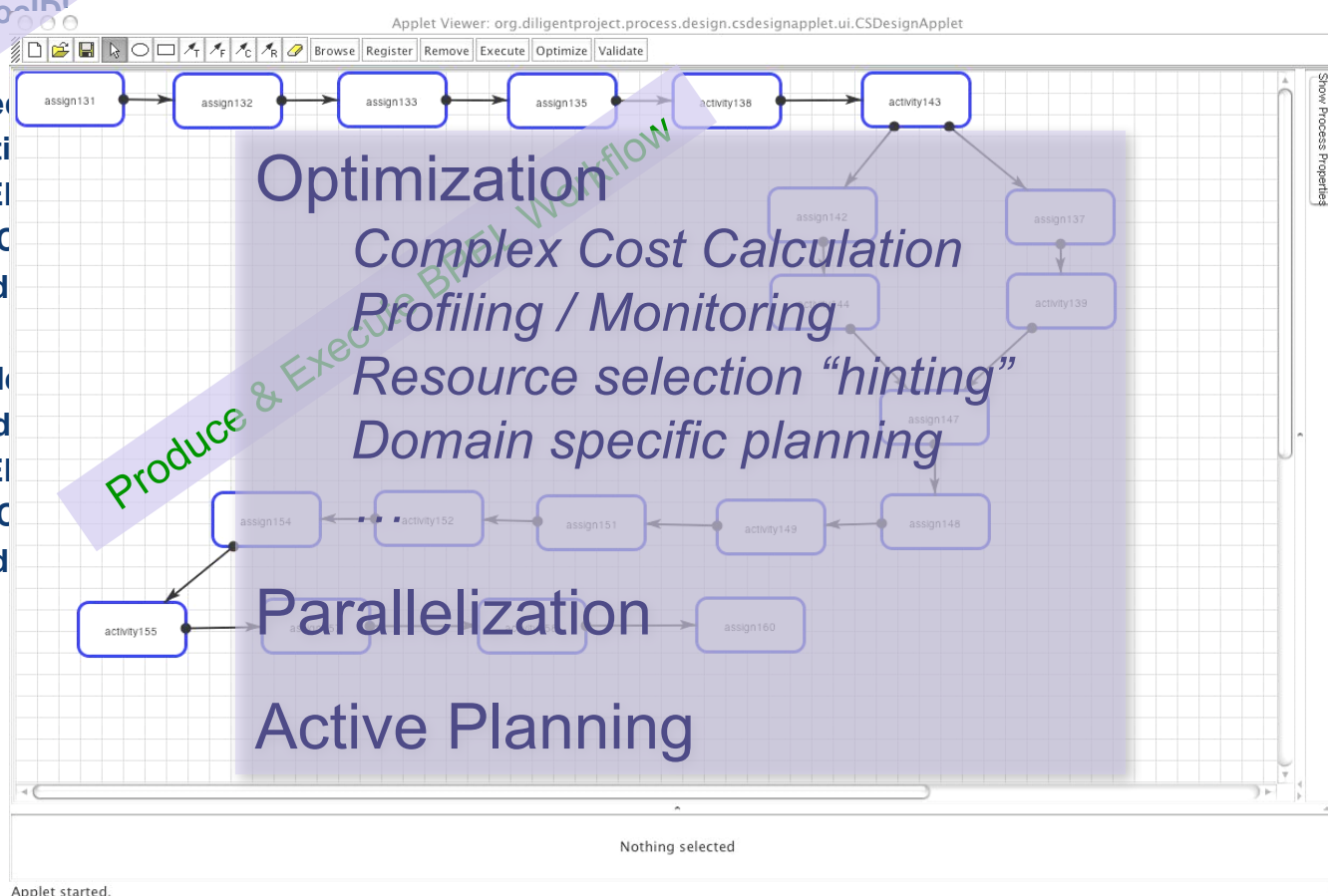


- Numerous Search services, for info retrieval & processing
 - Structured data and XML processing (scanners, sorters, joiners, filterers, transformers, retrievers)
 - Lookups (indices, FT indices, XML indices, Geo indices)
 - Content-based searches
 - External source probes
 - Fusion / Merging of results
- Query language (internal) for interfacing
- Workflow language (BPEL) for execution
- Data transport mechanism (ResultSet) for communication

project by 'title', 'description', 'subject'
on (keeptop 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**