d4science

D4Science-II
Preparation Meeting

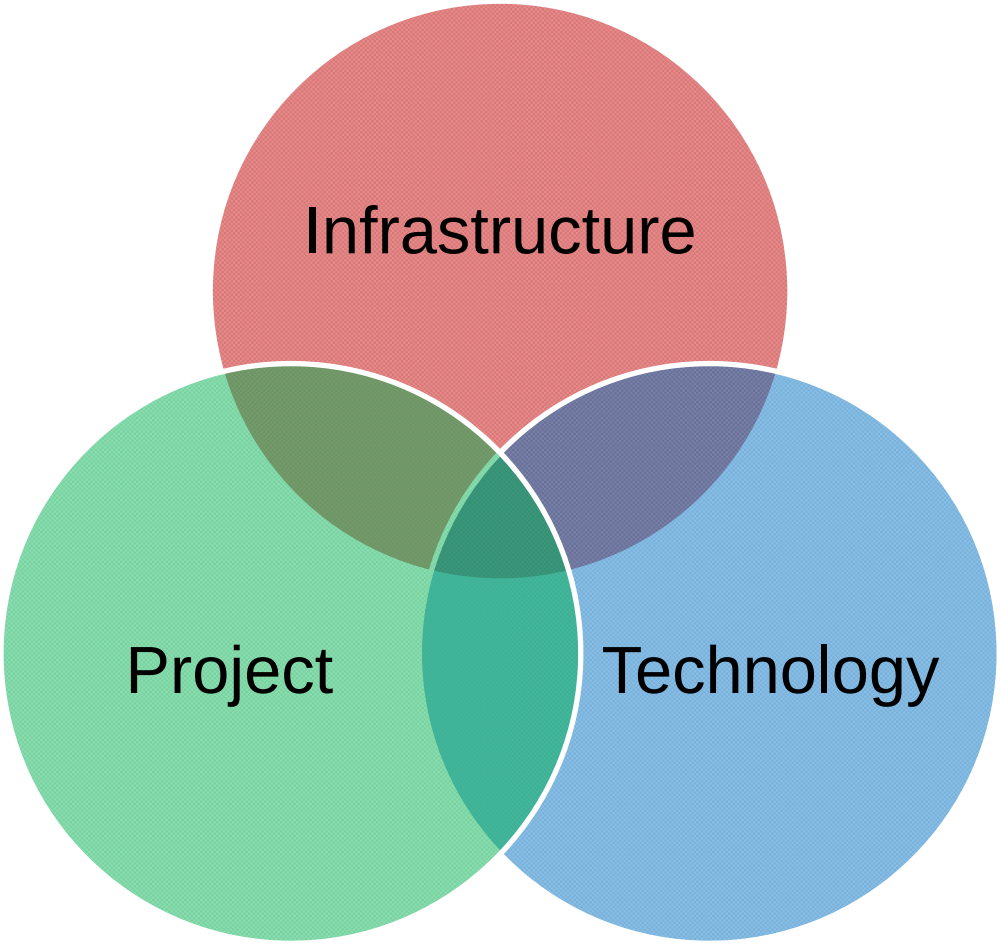
28th 2009
CERN (Geneva)

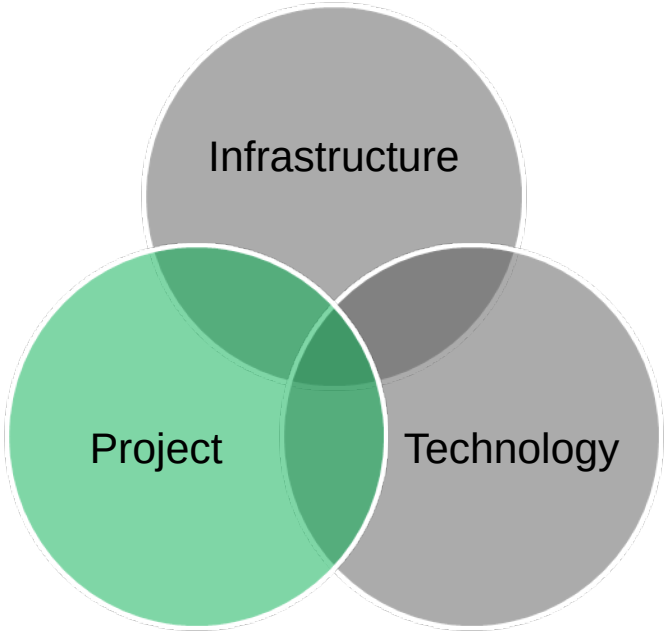
D4Science Objectives, Status, and Plans

Pedro Andrade pedro.andrade@cern.ch
Andrea Manzi andrea.manzi@cern.ch

e-infrastructure

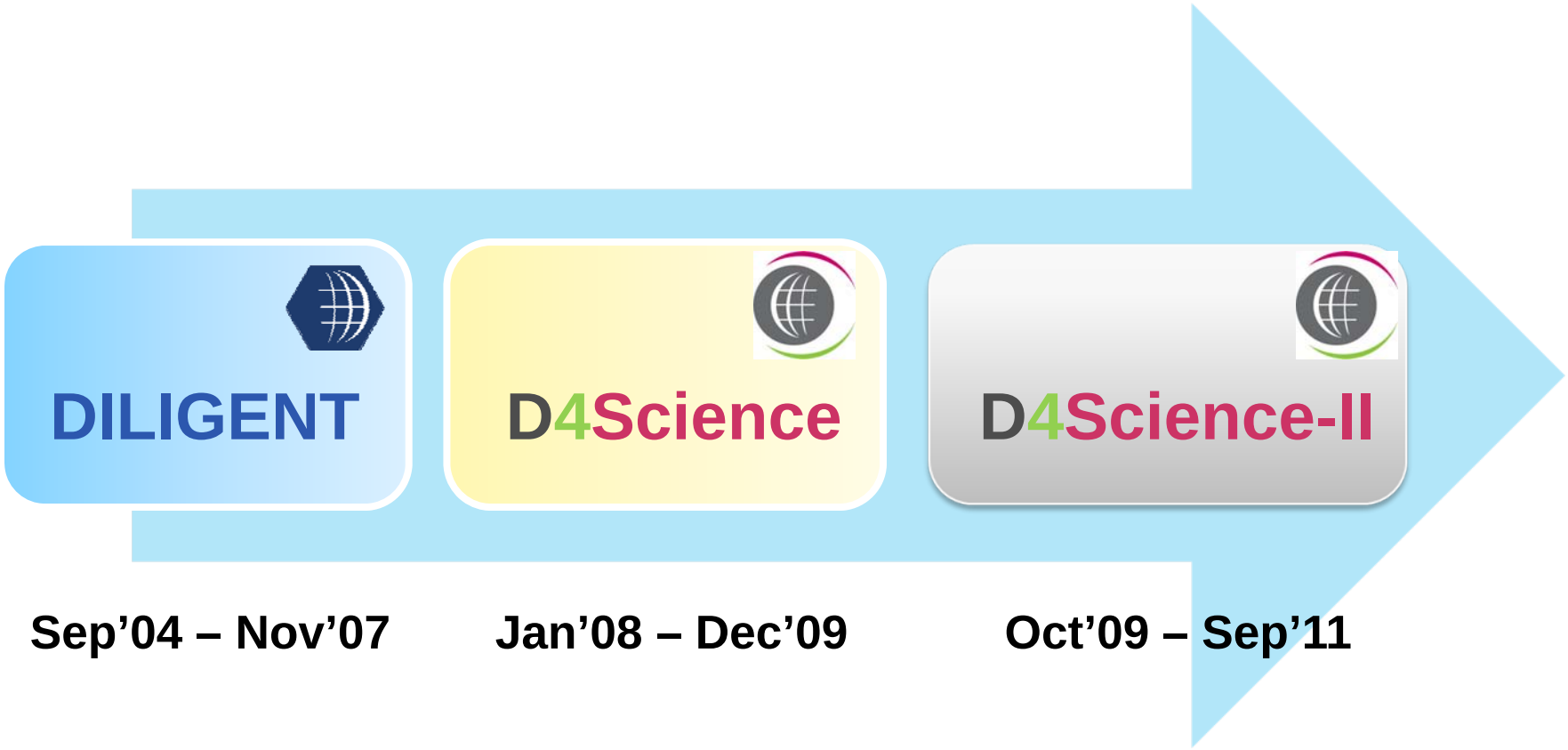
Outline





PROJECT

From DILIGENT to D4Science-II



From Sept 2004 to Nov 2007

- Digital Library e-infrastructure test-bed
- Based on EGEE infrastructure and middleware
- Cross-fertilization between the Grid and DL communities
- gCube prototype system
- e-Infrastructures as basic frameworks for creating DLs
- Two Virtual DLs:
  - ARTE (Humanities)
  - ImpECt (Environmental Monitoring)

D4Science Grand Objective

"The project will deploy, progressively consolidate and expand the e-Infrastructures built so far by the EGEE and DILIGENT projects so that they address the needs of several new scientific communities affiliated with the broad disciplines of **Environmental Monitoring (EM)** and **Fisheries and Aquaculture Resources Management (FARM)"**

D4Science Objectives

Test-bed Infrastructure

Production infrastructure

- procedures & policies
- system deployment & maintenance
- support to problem solving
- resource registration and monitoring

D4Science Objectives

VDLs

VREs serving EM and FARM

- from VDLs to VREs
- support to operative, cross-domain scientific communities
- porting/development of community specific resources

D4Science Objectives

gCube prototype

Consolidate and extend gCube

- better QoS
- infrastructure management
- VO, VRE definition and management
- information management, access & presentation
- collaboration supporting tools

D4Science Objectives

new DL basic framework

Strategies for exploitation, sustainability and outreach

- applying systematic dissemination & training
- liaising with other projects & initiatives
- reaching user scientific communities

Project Details

Full Name:

- D4Science > DILIGENT4Science
- **DI**stributed col**L**aboratories **I**nfrastructure on **G**rid **EN**abled **T**echnology **4 Science**

Activity Area: INFRA-2007-1.2.2 - Deployment of eInfrastructures for scientific communities

Project Type: Combination of Collaborative projects & Coordination and support actions (CCPCSA)

Budget: 3 916 735 € (3 150 000 € EC contribution)

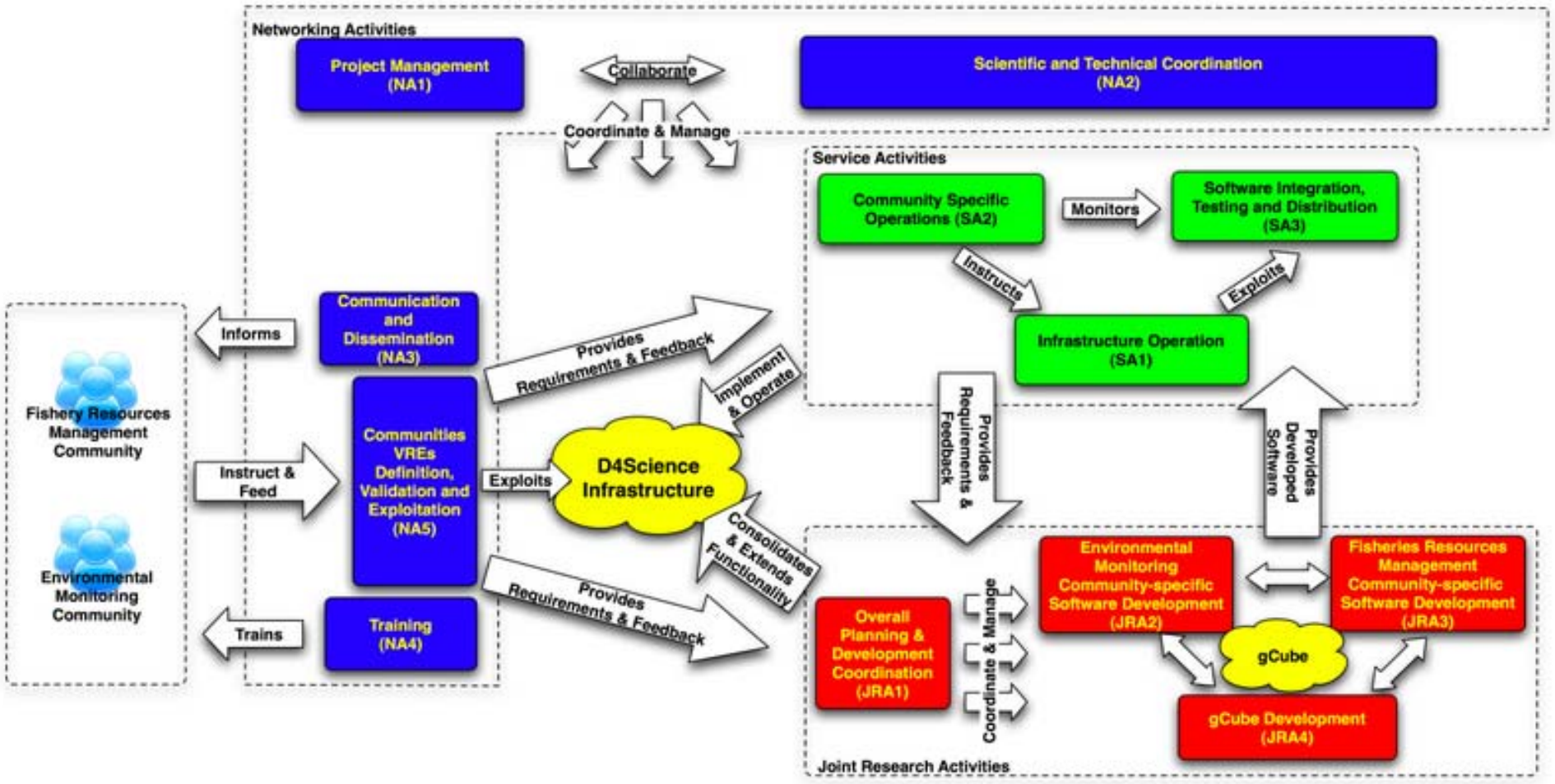
Effort: 400 p/m

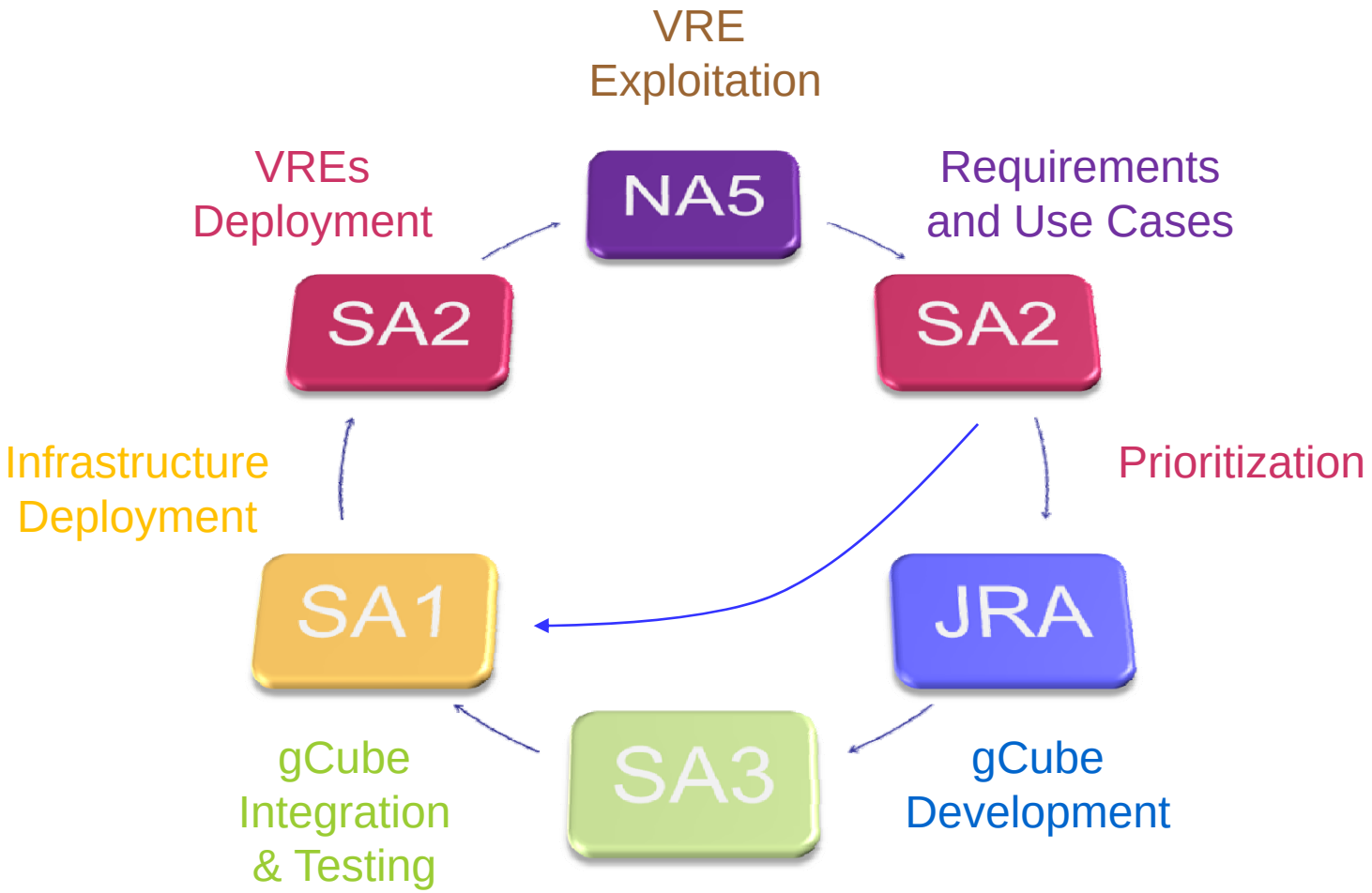
Duration: From Jan 2008 to Dec 2009 (24 months)

Partners

- ERCIM (FR)
- CNR-ISTI (IT)
- University of Athens (GR)
- CERN (CH)
- Engineering Ingegneria Informatica (IT)
- University of Strathclyde (UK)
- University of Basel (CH)
- European Space Agency (FR)
- UN Food and Agriculture Organization (IT)
- World Fish Center (MY)
- 4D SOFT (HU)

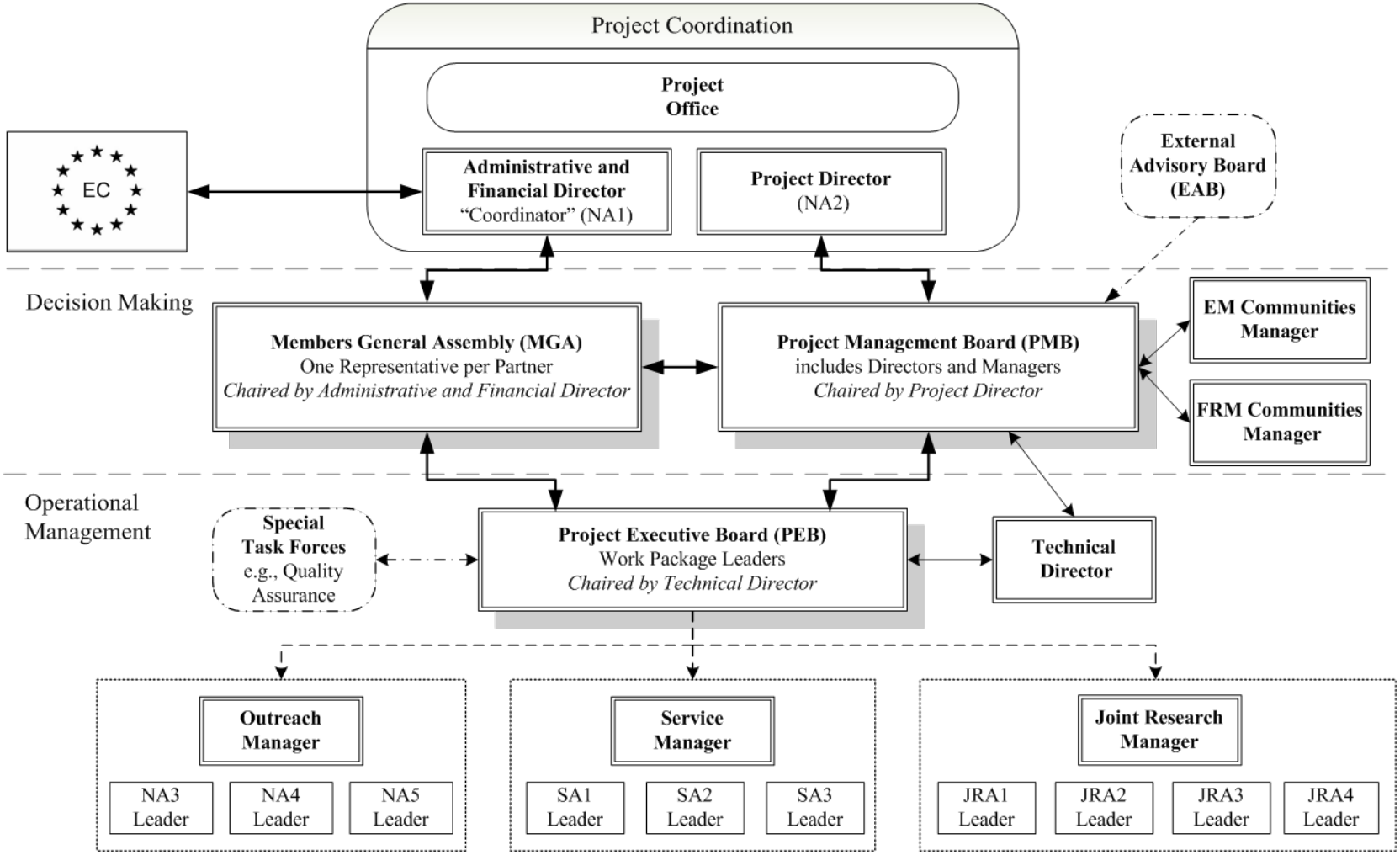
Activities


Networking Activities
Project Management (NA1)
Collaborate
Coordinate & Manage
Scientific and Technical Coordination (NA2)
Service Activities
Community Specific Operations (SA2)
Monitors
Software Integration, Testing and Distribution (SA3)
Instructs
Exploits
Infrastructure Operation (SA1)
Fishery Resources Management Community
Environmental Monitoring Community
Informs
Communication and Dissemination (NA3)
Provides Requirements & Feedback
Implement & Operate
Instruct & Feed
Communities VREs Definition, Validation and Exploitation (NA5)
Exploits
D4Science Infrastructure
Provides Requirements & Feedback
Provides Developed Software
Consolidates & Extends Functionality
Trains
Training (NA4)
Provides Requirements & Feedback
Overall Planning & Development Coordination (JRA1)
Coordinate & Manage
Environmental Monitoring Community-specific Software Development (JRA2)
Fisheries Resources Management Community-specific Software Development (JRA3)
gCube
gCube Development (JRA4)
Joint Research Activities

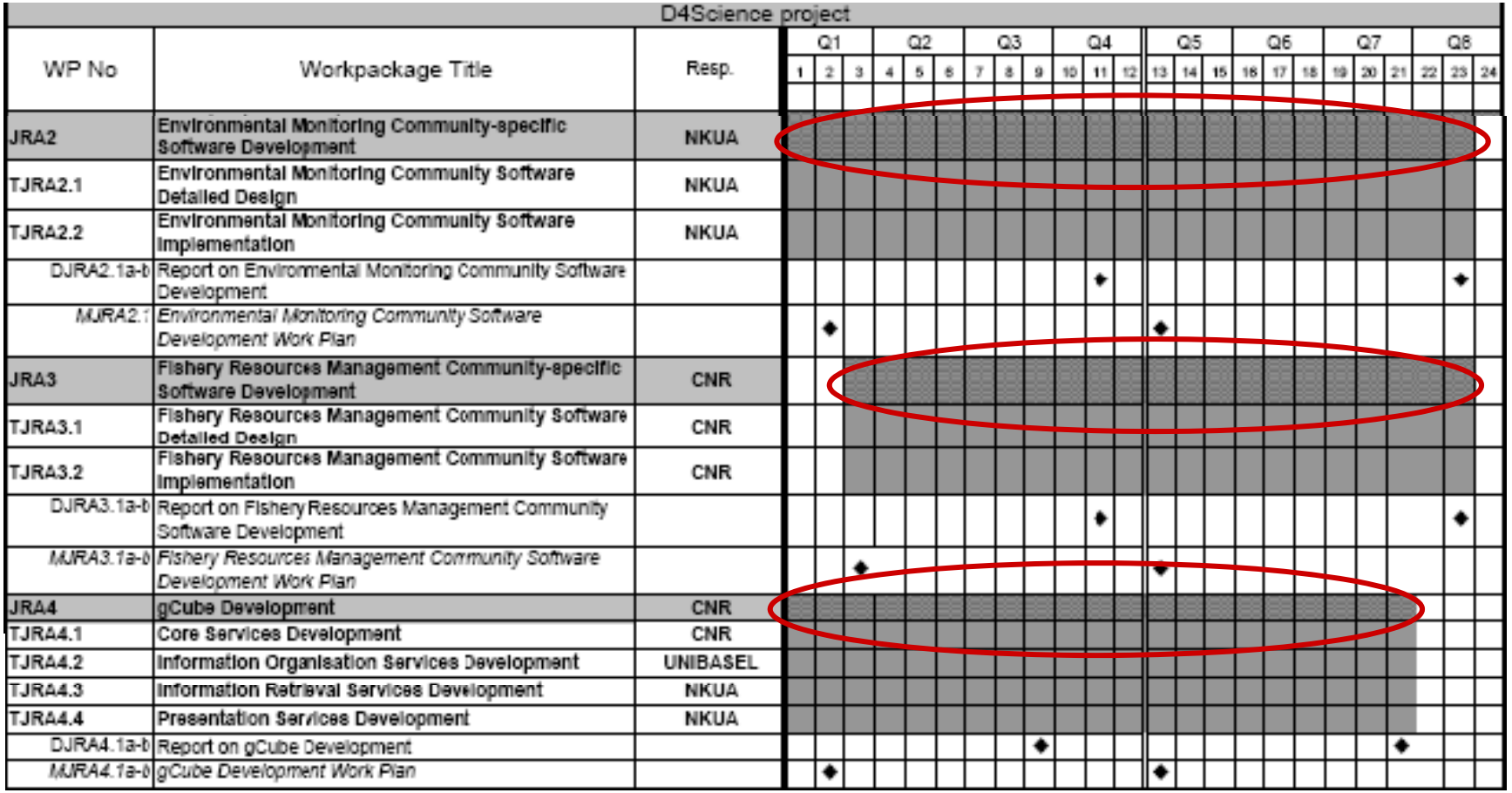

VRE
Exploitation
VREs
Deployment
NA5
Requirements
and Use Cases
SA2
SA2
Infrastructure
Deployment
Prioritization
SA1
JRA
SA3
gCube
Integration
& Testing
gCube
Development


Ticket
System
Code
Repository
Build and
Reporting
System
Software
Repository
gCube
Portal

Organisational Structure



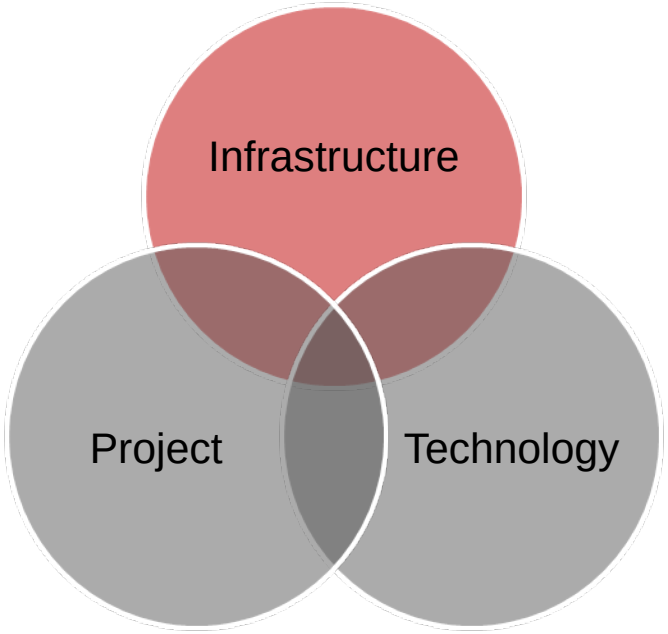
Major Milestones



Major Milestones

D4Science project			Q1			Q2			Q3			Q4			Q5			Q6			Q7			Q8		
WP No	Workpackage Title	Resp.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
SA1	Infrastructure Operation	CERN																								
TSA1.1	Procedures and resources planning	CERN																								
TSA1.2	Sites installation and validation	CNR																								
TSA1.3	Infrastructure monitoring	CERN																								
TSA1.4	Sites support	CNR																								
DSA1.1a-b	Procedures and resources plan		◆												◆											
DSA1.2a-c	Middleware deployment and operational support procedures					◆					◆							◆								
DSA1.3a-c	Infrastructure operation report								◆						◆											
MSA1.1	*Production Infrastructure at M6*							◆																		
MSA1.2	*Production Infrastructure at M11*												◆													
MSA1.3	*Production Infrastructure at M23*																								◆	
SA2	Community Specific Operations	CNR																								
TSA2.1	Environmental Monitoring Community resources operation	NKUA																								
TSA2.2	Fishery Resources Management Community resources operation	CNR																								
DSA2.1a-b	Resources Inventory and Plans					◆	ongoing								◆											
DSA2.2a-b	Community specific operation activity report													◆												◆
MSA2.1a-c	*D4Science Portal Release*							◆				◆												◆		

M6 M11 M23



INFRASTRUCTURE

Overview

D4Science vision: provide a scientific e-Infrastructure that removes heterogeneity, sustainability, scalability, and other technical concerns and enables its users to focus on their collaborative work.

- Easy-to-use tools for resources registration & management
- Cost-effective tools for data resource registration, metadata generation, and curation
- Seamless access to shared, heterogeneous, & distributed resources organized in dynamically created VREs
- Workspace for storing, enriching, annotating, and sharing data and compound objects to facilitate collaboration

A D4Science based infrastructure manages:

Hardware:

- Storage, Computing
- gCube Container

Services & Applications:

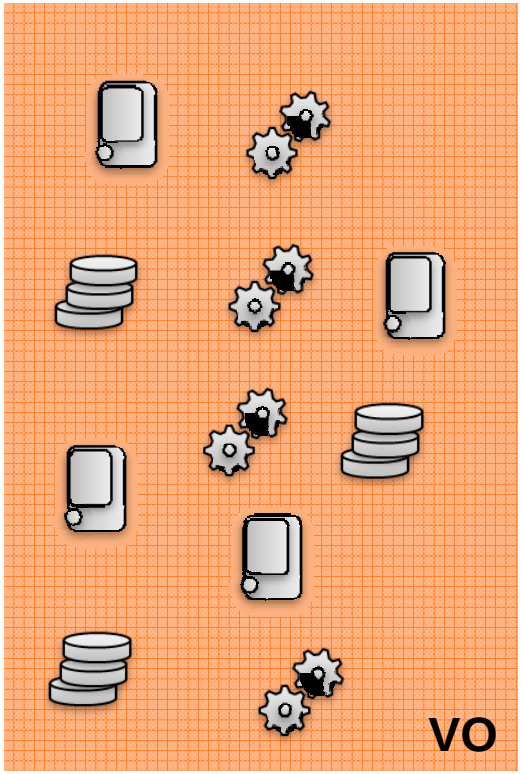
- gCube Web Services
- External Software

Collections & Auxiliary Resources:

- Data, Metadata, Indexes, Annotations
- Schemas, Mappings, Transformation programs

Virtual Organization

A **Virtual Organization** models sets of users and resources belonging to a e-Infrastructure.



It defines :

- What is shared
- Who is allowed to share
- The sharing conditions

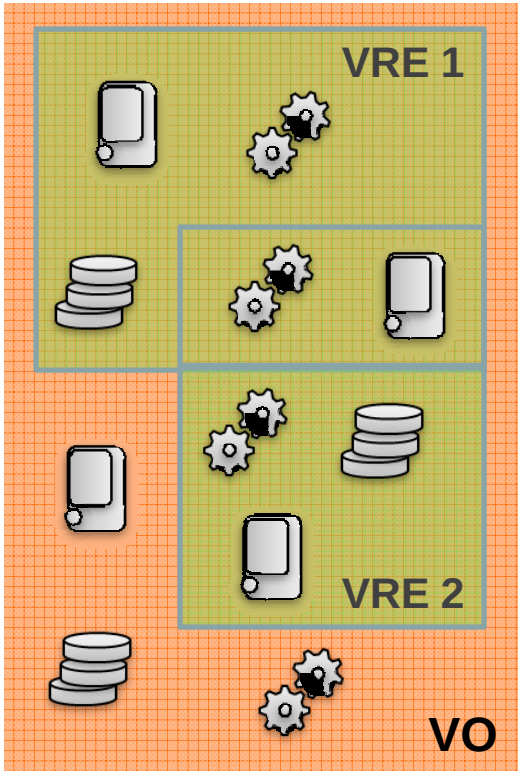
VOs:

- may have a limited lifetime
- may span multiple actual organizations

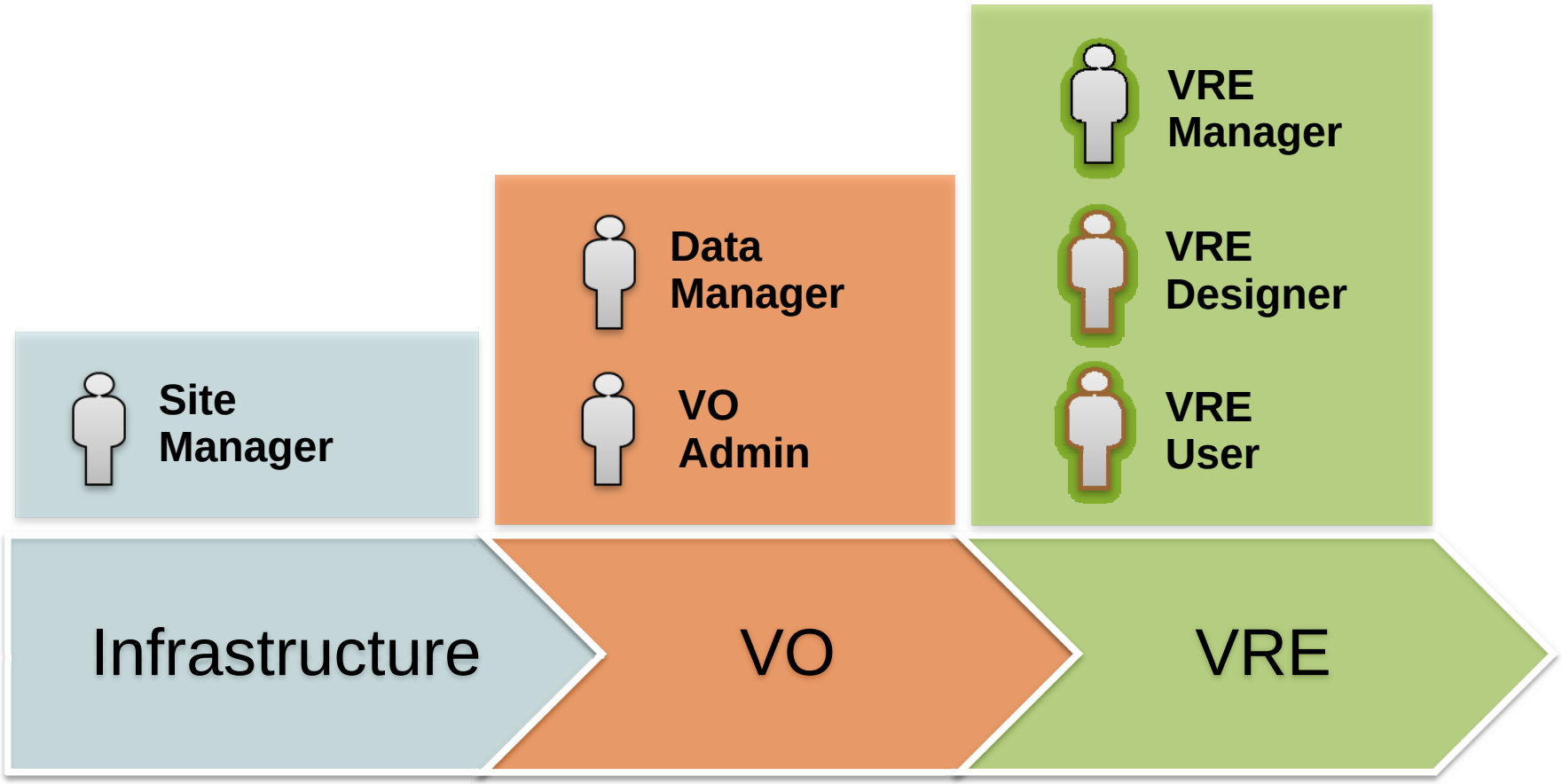
Virtual Research Environment

A **Virtual Research Environment** provides a framework of applications, services and data sources dynamically identified to support the processes of research/collaboration/cooperation.

The purpose of a VRE is to help selected VO members to carry out cooperative activities like data analysis and processing, and to produce new knowledge using specialized tools.

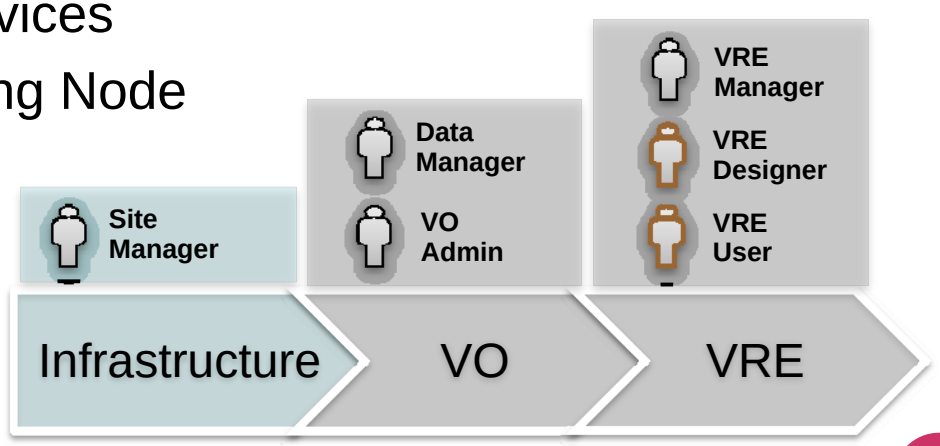


Infrastructure Roles



Infrastructure is based on hardware provided by the project members. Nodes and services are deployed and maintained by **Site Managers**. These nodes run:

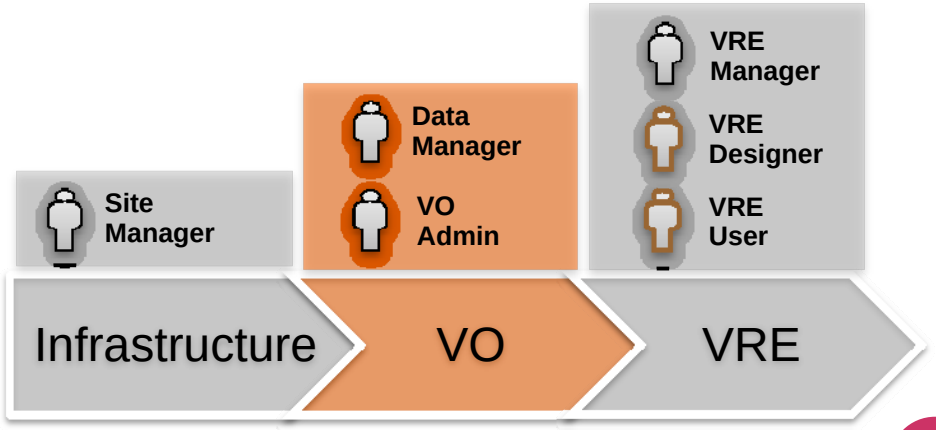
- gLite nodes
  - Running CE, SE, WMS, LFC, VOMS, MyProxy services
- gCube nodes
  - Running Infrastructure Enabling services
  - Running VO/VRE level services
  - Based on the gCube Hosting Node



Data collections staging by **Data Manager**

VOs creation and management by **VO-Admin**:

- Management of VO users and roles
- Deployment of VO services (Content, Metadata, Index, etc)
- Approval of VO resources



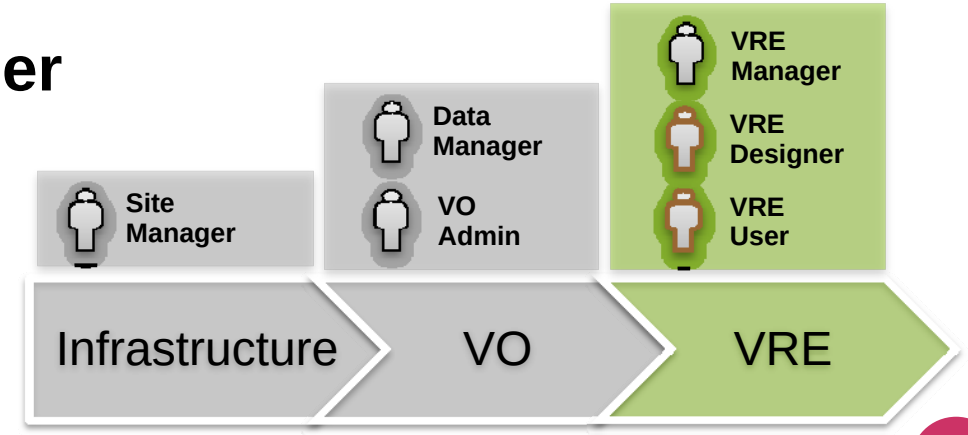
VRE creation and management by **VRE-Manager**:

- Management of VRE users and roles
- Deployment of VRE services (Annotation, Search, etc)

VRE definition by **VRE-Designer**:

- Selection of data collections and functionality for VRE

VRE exploitation by **VRE-User**



Operational Procedures

Infrastructure operation is based on a well established set of procedures covering many infrastructure activities:

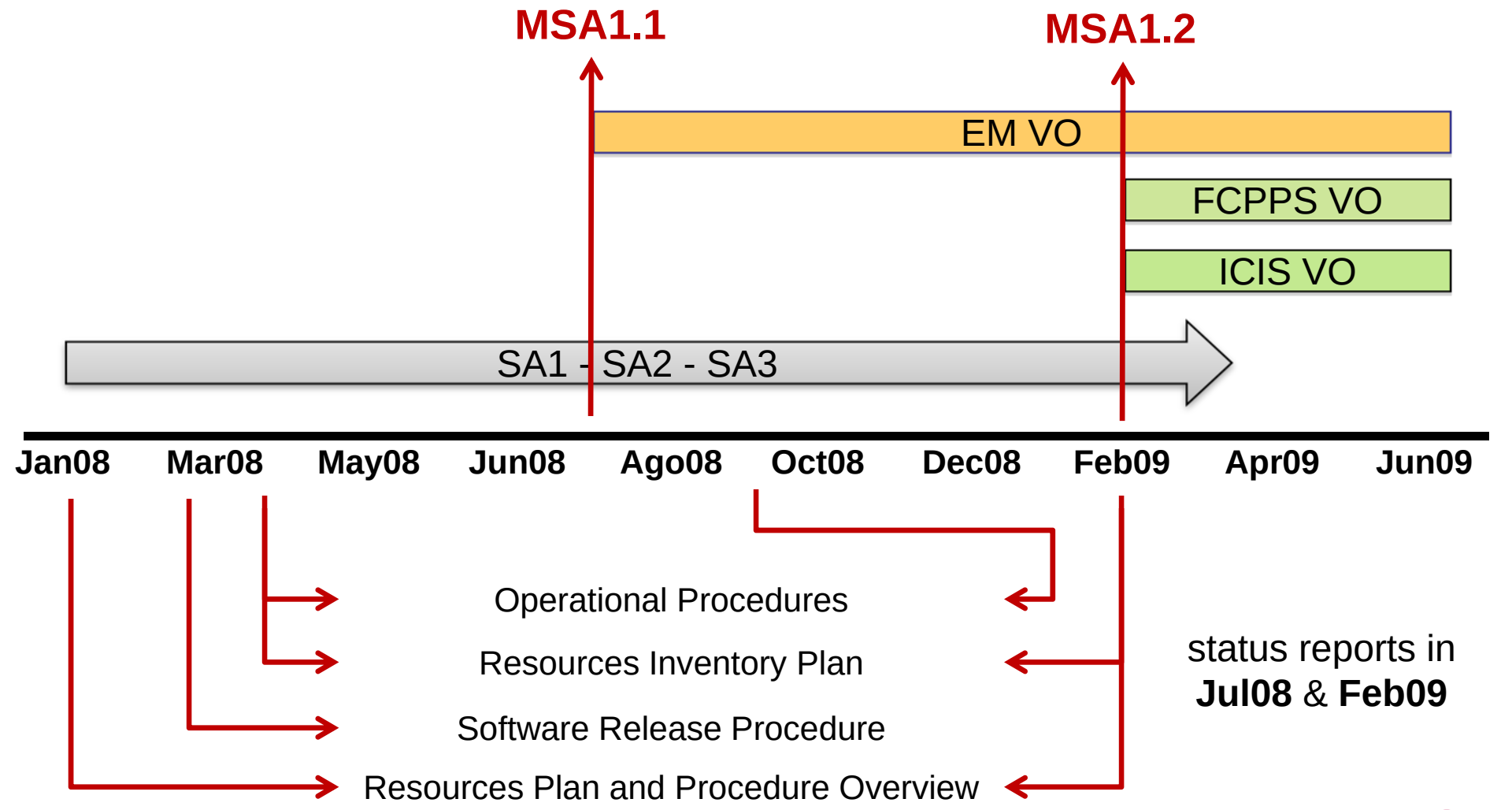
- **Software**: gCube Release Cycle
- **Nodes**: Deployment, Certification, Downtime
- **Infrastructure:** Monitoring, Security
- **Data**: Collection Validation
- **Support**: Incident Management, Service Request
- **VRE**: Deployment, Validation, UC Resources, Access

Described in project deliverables and infrastructure website

Supported by many **collaboration tools** (ETICS, TRAC, Wiki)

Consolidation and extension of EGEE/DILIGENT procedures

Infrastructure Status



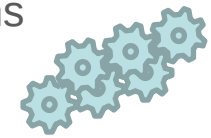
Infrastructure Status

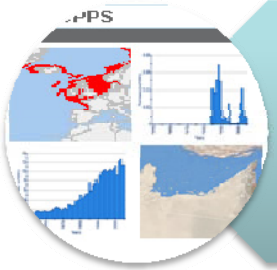
EM VO

- Global Vegetation Monitoring VRE
- Global Chlorophyll Monitoring VRE
- Demo VRE


41 gHNs


22 Collections


197 RIs

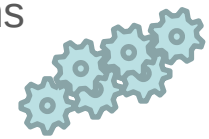


FARM VO

- Fisheries Country Profiles VRE
- Demo VRE


36 gHNs


11 Collections

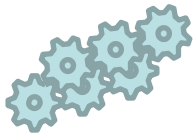

169 RIs

ICIS VO

- AquaMaps VRE


35 gHNs


5 Collections


148 RIs

Infrastructure Status

Infrastructure:

- 4 site providers: CNR, ESA, NKUA, UNIBASEL
- 1 central coordinator: CERN
- Running gLite 3.1 (EGEE) & gCube 1.1.6 (D4Science SA3)
- Part of EGEE production infrastructure

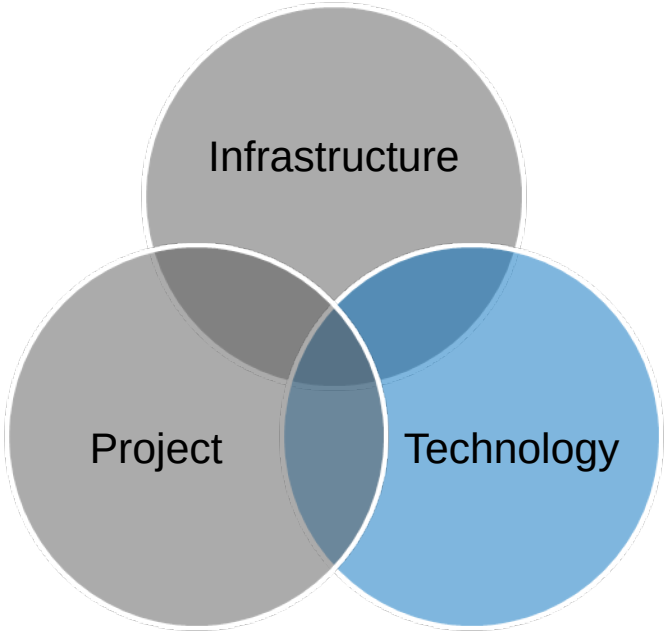
gCube System:

- 2 Release Cycles: org.gcube.1.0 & org.gcube.1.1
- Packaging of 290 components in 82 software archives
- Build & Test Support Tools: ETICS, Report Tool, Analyzers
- Documentation: 3 manuals (dev, user, admin) and javadoc
- Distribution Site: dev, integration, and stable releases

From the experience deploying VOs/VREs for EM & FARM:

	ACTION	ROLE	TOOL	TIME
VO	deploy gHNs	Site Manager	CLI	1 min
	stage data collections	Data Manager	Portal	hours
	register resources	Resource Owner	Portal	10 min
	deploy VO services	VO Admin	Portal	30 min
	manage VO users and roles	VO Admin	Portal	5 min
VRE	define VRE functionality and data	VRE Designer	Portal	5 min
	deploy VRE services and layout	VRE Manager	Portal	10 min
	manage VRE users and roles	VRE Manager	Portal	5 min

New VREs can be created in **few minutes**



TECHNOLOGY

gCube System:

- A system to enable **e-infrastructures** for the creation, hosting and maintenance of dynamic virtual environments capable of satisfying research and collaboration needs of distributed scientific communities organized in VOs
- Lifts the Grid approach of batch jobs to Web Services deployment and invocation in a SOA e-Infrastructure

gCube Core Framework:

- Standardizes all systemic aspects of service development;
- Promotes the adoption of best practices in multiprogramming and distributed programming

The gCube Core Framework (gCF)

- An application framework for the consolidation and development of existing or new gCube services
- Reduces the complexity in the design and implementation of gCube services:

The gCube Core Distribution (gCore)

- An easy-to-install, self-contained sandbox to participate to D4Science e-Infrastructure
- Simplifies the tasks of system administrators, infrastructure managers, and resource providers:

gCube Numbers

gCube is inherently complex & large due to

- The vast functional domain it covers
- The required abstractions allowing collaborative development and openness

Distinguished build elements as of release v1.1.6:

- 65 Services and associated libraries
- 35 Portlets & servlets
- 54 Distribution packages
- Corresponding testsuites, service stubs, archives, …

Code size:

- Packages: 799, Classes: 4.406, Methods: 30.285, NCSS: 305.039

Building blocks' characteristics:

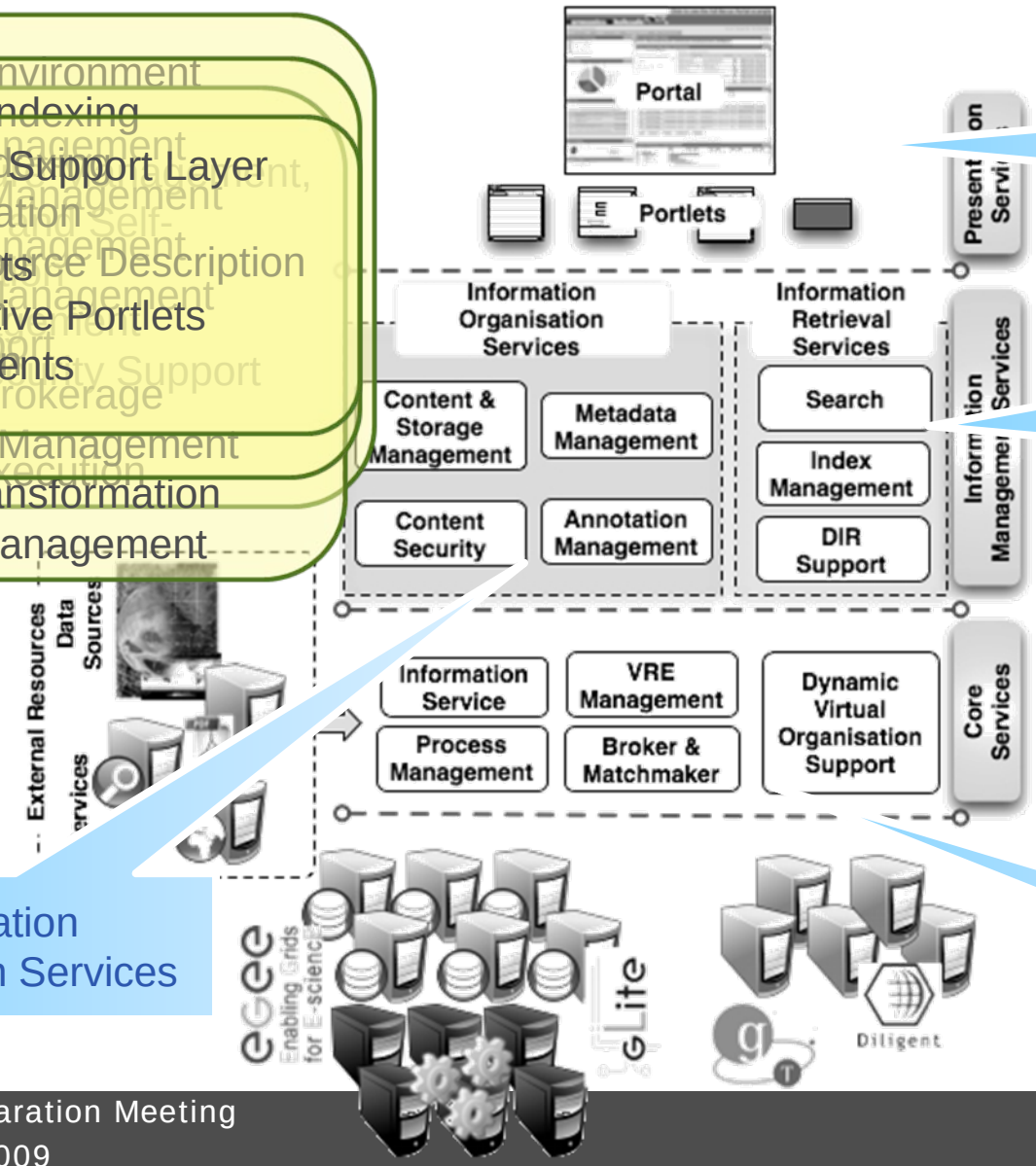
- Highly sophisticated, composite sub-systems (accounting for more than 80%)
- 3 large frameworks (for building higher level elements)

Implementation team size:

- Constantly more than 20 developers & designers

gCube Architecture

- Runtime Environment
- Metadata Indexing
- Storage Management
- Application Support Layer
- Collection Management
- Portals
- Content Management
- User Portlets
- Metadata Management
- Administrative Portlets
- Archive Import
- Desktop clients
- Metadata Brokerage
- Search
- Annotation Management
- Process Execution
- Content Transformation
- Ontology Management



Specifications, Standards & Technologies

- WS-*, WSRF, WS-BPEL
- JSR
- Glue Schema
- GSI-Security

- Java
- Globus Toolkit
- gLite

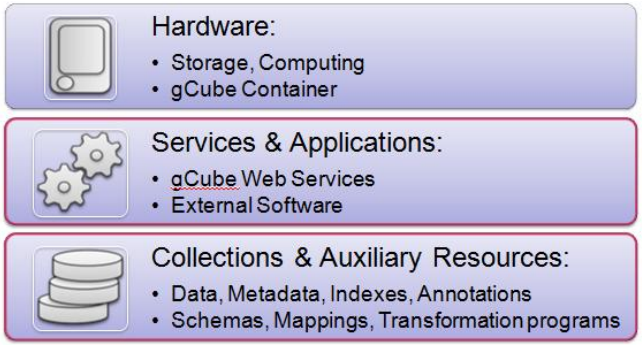
- Distributed under Open Source License: EUPL

More Exploited:
- DC
- ISO

More coming:
- OAI-PMH & OAI-ORE
- WS-DAI
- OpenSearch
- OpenGIS - related

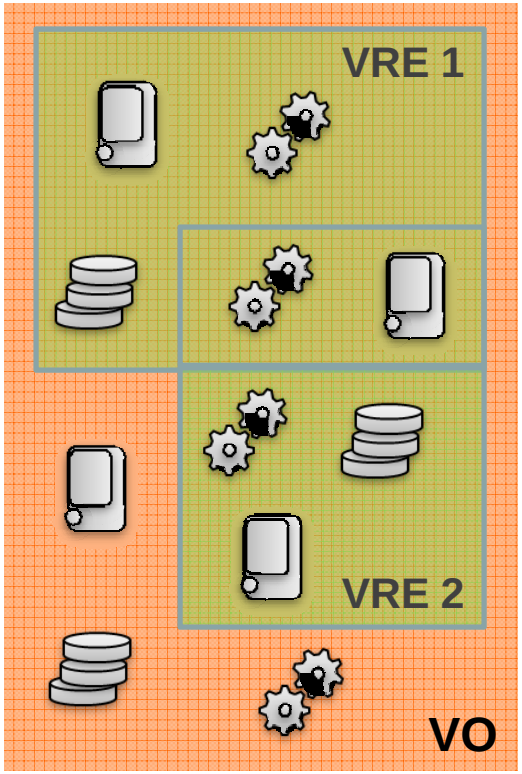
gCube **Information System** collects information about the capabilities and status of all resources:

- Glue schema for computational and storage resources
- Profiles for gCube services and their running instances
- Profiles for content and metadata collections
- Allows new resource types to be registered
- Allows extending resource descriptors with arbitrary information
- Is exploitable through plain standards
  - XML, XQuery
- Implements Resource Scoping
  - Infrastructure, VO, and VRE



gCube **VRE Management** System:

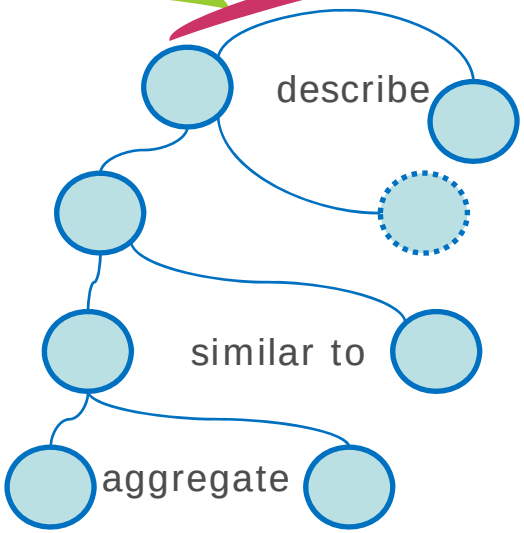
- Manages services and applications
- Reduces deployment costs
- Grants execution only to certified SW
- Dynamic Deployment of Services
- Automatic service reminiscence
- Dynamic reconfiguration:
  - Triggered through monitoring
  - Dynamic deployment
  - Service reminiscence capacities

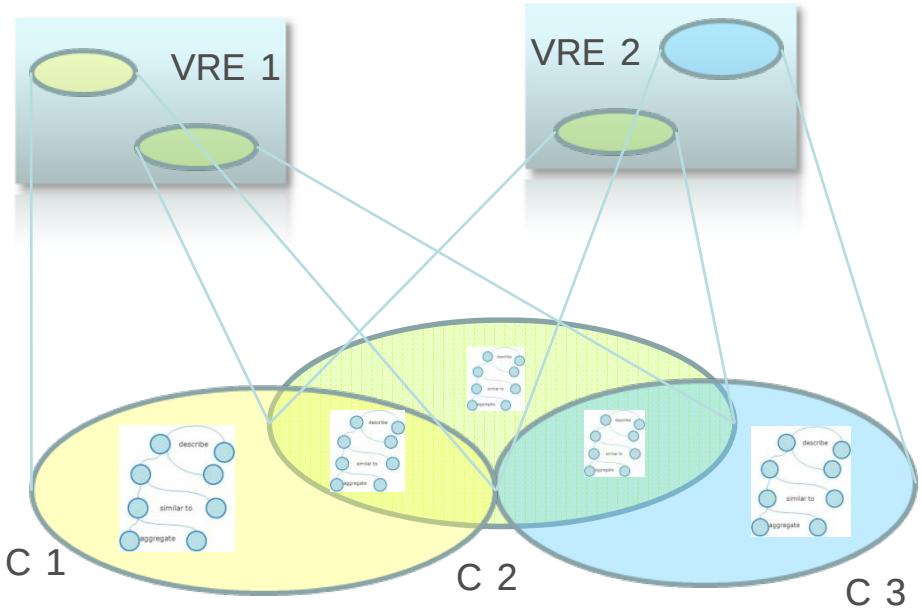


gCube Functionality

gCube **Data Management** System:

- Persistently stores compound objects
- Manages heterogeneous metadata
- Supports metadata cleaning, enrichment, and transformation by exploiting mapping schema, controlled vocabulary, thesauri, and ontology

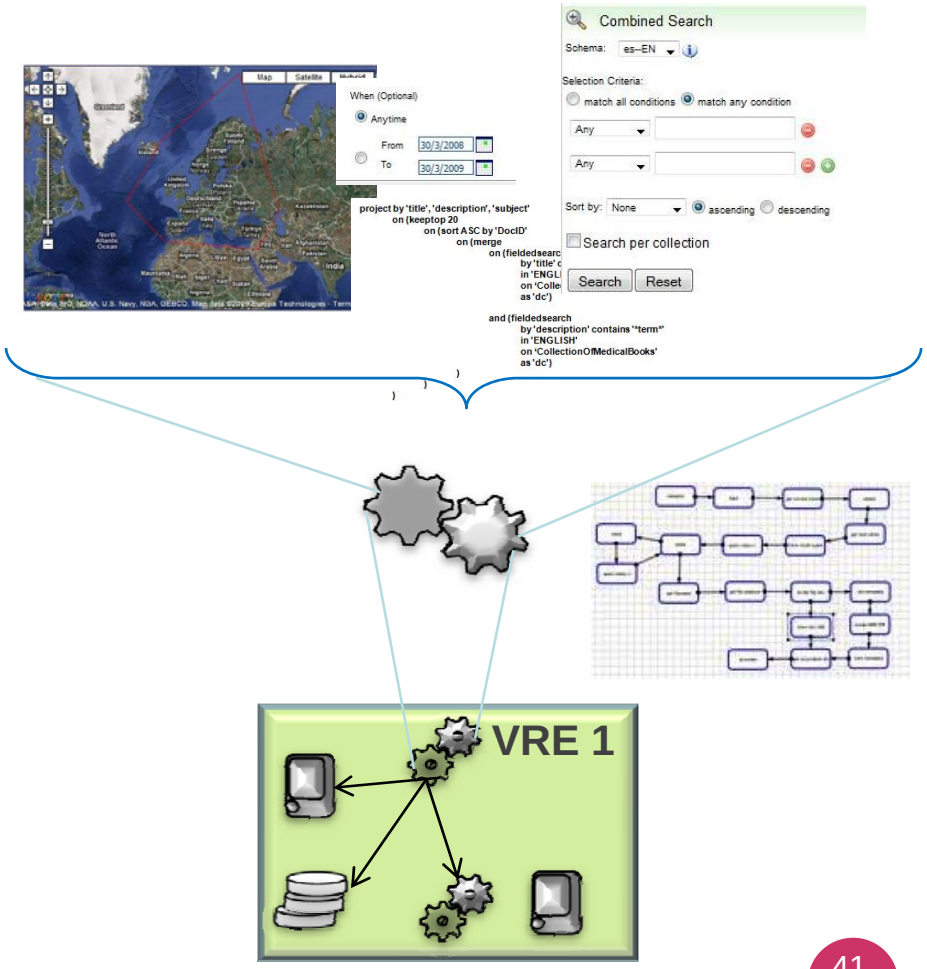




- Supports programmatic/manual annotation of content, e.g. data provenance
- Supports content linking
- Provides support for collections
- Supports collections sharing across VREs

gCube **Search Management:**

- provides an XML-based query language over full text, geospatial, and temporal information
- Maximizes the resources available to VRE users by promoting resource sharing and avoiding suboptimal usage
- Combines information retrieval and data processing capabilities

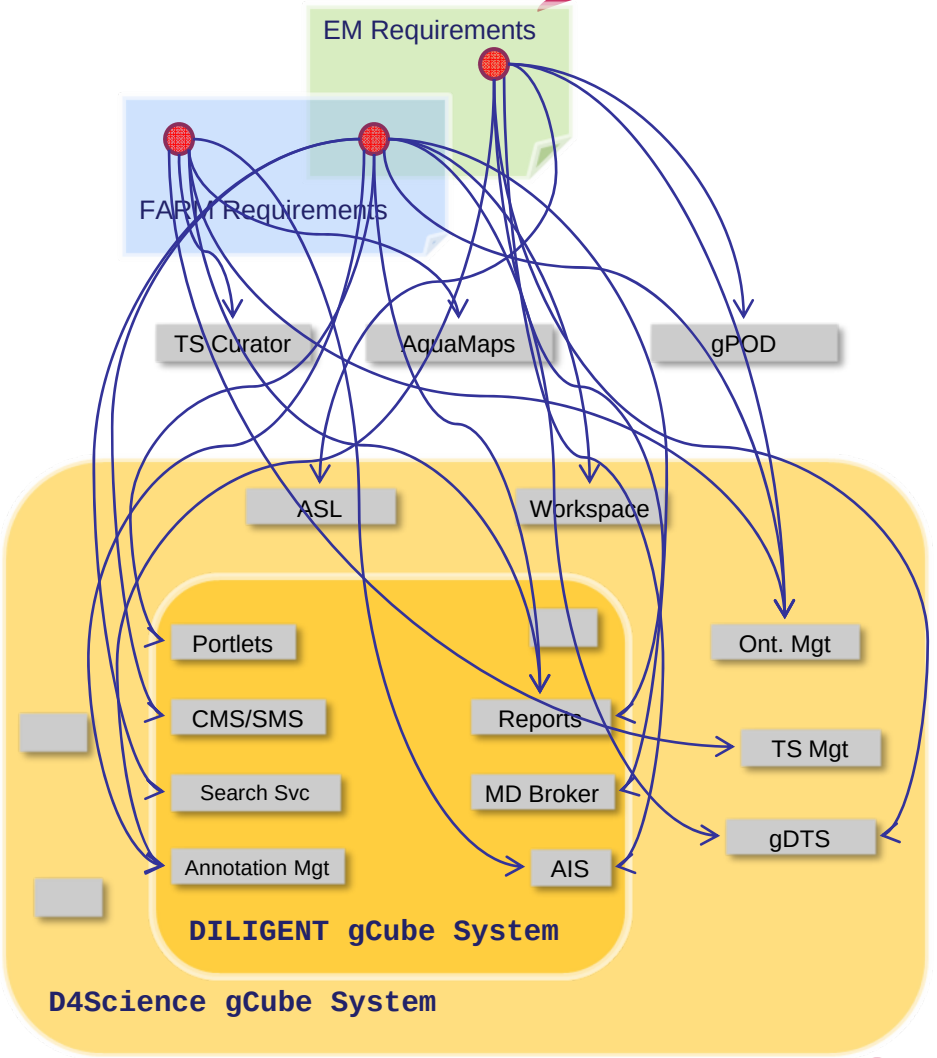


The gCube **Search Management** exploits a data-oriented processing analogous to HPC computation facilities

- Supports dynamic inclusion of arbitrary non-native data and logic providers
  - Ad-hoc processing of information at any stage: Operators
  - Defines and exploits semantic information of services
- Exploits a dataflow execution engine
- The gRS formalizes the exchange of large data sets in web services (paging, store & forward, throttling / flow control…)
- Adds the “by-ref” notion to data exchanged via services
- Confronts several performance issues of WS interactions

Community Requirements

- Specific workpackages to deal with user community requirements (JRA2/3)
- Common requirements are included in the baseline gCube system
- Yet, some remain community specific:
  - FARM AquaMaps
  - FARM TS curator
  - EM gPod integration



Production Portal:

- http://portal.d4science.research-infrastructures.eu

gCube Infrastructure Monitoring

- http://monitor.d4science.research-infrastructures.eu

Thanks !!!

Questions ???

http://www.d4science.eu/resources

Community

Researchers and stakeholders operating over a widespread geographic scale to provide political and technological solutions to global environmental issues (e.g., marine environment, forest ecosystem, air quality)

Requirements

Secure Virtual Research Environments where access to huge amount of information, both products or different kinds of reports, added-value applications and services, definition of workflows and on-demand processing of data are all seamless tasks

Community

Worldwide spread researchers and decision-makers from many disciplines (biologists, climatologists, GIS experts, socio-economists, fishery managers, etc.) operating to facilitate and secure the long term sustainable development and utilization of the world's fisheries and aquaculture

Requirements

VREs, encompassing many heterogeneous resources on aquatic biodiversity and socio-economics, offering to the communities tools for collaboration on shared fishery assessments in a continual way, instead of sporadically as at present

Deployment and Exploitation