Hybrid Data Infrastructures: concept, technology and the complex ENVRI RIs use case

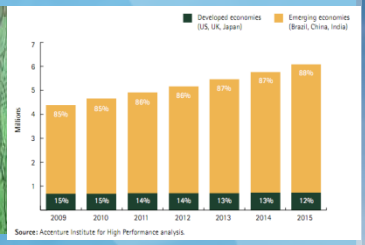
Pasquale Pagano
Italian National Research Council (ISTI-CNR)
pasquale.pagano@isti.cnr.it

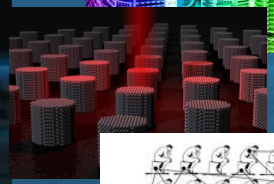
New Science Pattern

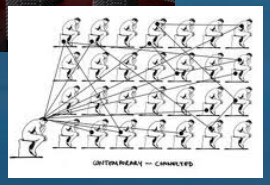
The Context

Science is increasingly *global*, *multipolar*, and *networked*

Data continue to grow in *Volume*, *Variety*, and collection, processing and consumption *Velocity*







The Needs

Computational environments dealing with the volume of the data

Efficient and tailored **storage and access technologies** dealing with the variety of the data types

Elastic management of the resources dealing with the innovative approaches for collection, processing and consumption of the data

World-wide collaborative environment between distributed scientific communities dealing with the federation of heterogeneous data sources

The Solution

Hybrid Data Infrastructures

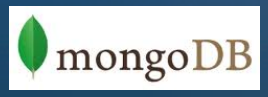
integrated technologies supporting efficient data management

D4Science Hybrid Data Infrastructure

- Availability of typical biodiversity processes running on **computational and storage resources** offered by grid and cloud resource providers
- New technologies generally identified as **no-sql databases** as service
- Accessibility of distributed computing platform supporting **MapReduce**
- Porting to MapReduce of several algorithms for performing **data analysis and mining**
- **Geographical data management** support

D4Science HDI hosts biodiversity communities federated by the iMarine and the EUBrazilOpenBio initiatives

D4Science HDI will provide ENVRI RIs with seed resources



D4Science Technology: the gCube system

- **gCube** offers solutions to **abstract** over differences in location, protocols, and models by
  - scaling no less than the interfaced resources,
  - keeping failures partial and temporary,
  - reacting and recovering from a large number of potential issues.
- gCube doesn't hide infrastructures middleware and technologies.
- gCube turns infrastructures and technologies into a **utility** by offering a single registration, monitoring, and access facilities.

gCube: Policy-oriented Security Facilities

Service Oriented Authorization, Authentication and Accounting (**SOA3**) is a security framework providing *security services* as web services, according to *Security as a Service* (**SecaaS**) research topic.

- Security as a Service
  - Authentication and Authorization provided by web services called by resource management modules
- Flexible authentication model
  - the user is not requested to have personal digital certificates
- Attribute-based Access Control
  - a generic way to manage access: access control decisions are based on one or more *attributes*
  - user related attributes (e.g. roles, groups) and environment related attributes (e.g. time, date)

https://gcube.wiki.gcube-system.org/gcube/index.php/Data_e-Infrastructure_Policy-oriented_Security_Facilities

gCube: Policy-oriented Security Facilities [cont.]

SOA3 Authentication Module provides *Authentication as a Service*.

The module receives *authentication requests*, matches received information with an external identities repository and returns the response as SAML assertion.

- Flexible authentication model
  - SOA3 provides a native authentication model based on userid/ password: X509 certificate based authentication is also supported
- RESTful interface
  - decouples the module from the underlying infrastructure according to the *zero-dependencies* model. Anyway the module is also usable as Java Library Based
- Based on SAML
  - user attributes are inserted in a standard SAML Assertion

gCube: Policy-oriented Security Facilities [cont.]

SOA3 Authorization Module bases its decisions on stored *policies* by which it is able to determine if a *subject* can perform a certain *action* on a certain *resource*.

The subject is defined by a set of *attributes* referred to caller, call and environment: **Attribute Based Access Control** (**ABAC**)

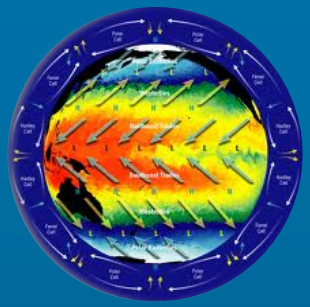
- Attribute-based Access Control Model
  - a generic and extensible model which base the decisions on a set of attributes from caller, call, and environment
- Policy-driven decisions
  - the decisions are based on a set of pre-defined stored policies
- Standard based architecture
  - the module is based on the standard *eXtensible Access Control Markup Language* (**XACML**) 2.0
- Extensible set of attributes
  - possibility to configure the attributes set to be used for policies evaluation

ENVRI: Environmental Research Infrastuctures

ENVRI	Title: Common Operations of Environmental Research Infrastructures
	Call Identifier: FP7-INFRASTRUCTURES-2011-1
	Starting Date: 01/11/2011
	Duration: 36 Months
	Keywords: Environmental Research Infrastructures Data processing, Interoperability, Reuse, GEOSS

Environmental Science

oceanic and atmospheric processes



long-term development of the climate system

biodiversity

development of the cryosphere and lithosphere

Earth as a single complex and coupled system

ENVRI Partners

Universities

- University of Amsterdam
- University of Helsinki
- Cardiff University
- University of Edinburgh
- University of Bremen

Research Centers

- Italian National Research Council (CNR)
- Centre National de la Recherche Scientifique (CNRS)
- Istituto Nazionale di Geofisica e Vulcanologia (INGV)
- Koninklijk Nederlands Meteorologissch Instituut
- Institut Francais de Recherche pour l'exploitation de la mer (IFREMER)

Agencies

- CEA- Commissariat à l'energie atomoque te aux ènergies alternatives
- ESA – European Space Agency
- EAA – Environment Agency Austria

Others

- CSC – Tieteen Tietotekniikan Keskus Oy Ltd.
- EISCAT Scientific Association
- EGI – European Grid Initiative

Goal

Enable multidisciplinary scientists to access, study and correlate data from multiple domains for “system level” research

by providing solutions and guidelines for the RIs common needs

ESFRI Environmental Research Infrastructures

- Tropospheric research aircraft

COPAL

- Upgrade of incoherent SCATter facility

EISCAT-3D

- Multidisciplinary seafloor observatory

EMSO

- Plate observing system

EPOS

- Global ocean observing infrastructure

EURO-ARGO

- Aircraft for global observing system

IAGOS

- Integrated carbon observation system

ICOS

- Biodiversity and ecosystem research infra

LIFEWATCH

- Svalbard arctic Earth observing system

SIOS

ESFRI Environmental RIs: complex infrastructure

Data acquisition is continuous

- Datasets are not static since data are continuously streamed from data sources
- Need a persistent identifier

Data stored in multiple sites

- Each site combines data from sources in different ways
- Not true replication
- Same data stream stored at different sites has a different persistent ID

Federated AAI

- Each site is responsible for authentication and authorization
- Common LDAP for users' credential with Shibboleth on top

Different access rights

- Anonymous for public data
- Read-only for not-public data
- Not-public data may become public after the embargo period is expired

RIs' Heritage

A variety of data

- complex and sometimes fuzzy
- heterogeneous and distributed
- primary and processed data

Existing practices

- data acquisition, validation and staging policies
- data consumption

Analytical and modeling platforms

- data driven methodologies
- data exchange and integration
- e-Laboratories

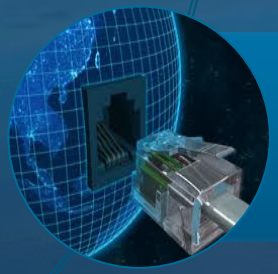
Technical Foundations

Standards and Recommendations

- INSPIRE Directive 2007/2/EC on environmental data sharing infrastructure
- Open Geospatial Consortium (OGC W*S) standards

e-Infrastructures

- EGI, D4Science
- GENESI-DEC, iMarine, EUDAT, …



Technologies

- Hadoop Map/Reduce
- NoSql storage solutions

Approach

PROVIDE SOFTWARE TOOLS TO

Promote Accessibility

discover data which are heterogeneous in format, content, and metadata description

harmonise, **integrate** and **analyse data** across domains and RIs

Preserve Specificity

Data Discovery and Access

Metadata Model

- Core set plus customisable attributes
- Compliant with INSPIRE Implementation Rules for Metadata

Tools

- Metadata Catalogue Services (OGC OpenSearch, CSW)
- Specific Gateways (to connect existing solutions not compliant with the adopted specifications)
- OGC Web Coverage Service to extract spatial subset of data

Outreach

- Register relevant components in GEOSS to interoperate with GEO-GEOSS
- Register data resources in the GEOSS Common Infrastructure

Data Integration, Harmonization, Analysis and Publication

Approach

- Exploit computational and storage capabilities of existing e-Infrastructures

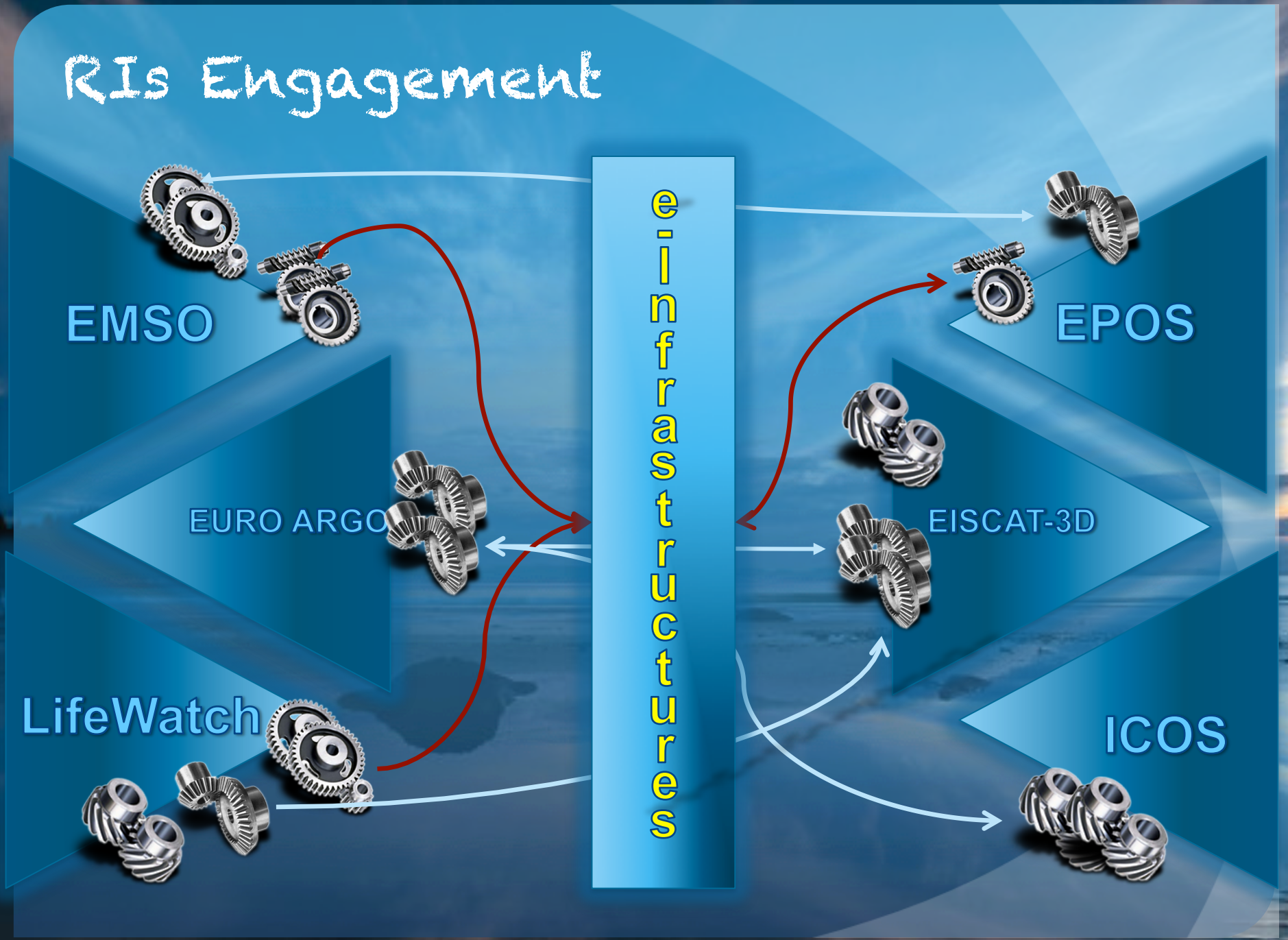
Tools

- Enable integration and harmonization
- Frameworks + plugins supporting temporal and spatial analysis

Outreach

- Linked Data for publishing and connecting structured data with non-collaborative consumers
- RDF and OWL to describe relations between e-Infrastructures components

RIs Engagement



Prototype: from discovery to process and publication

Discovery
- OpenSearch (OGC CSW 3.0)
- Federation of Catalogue Services

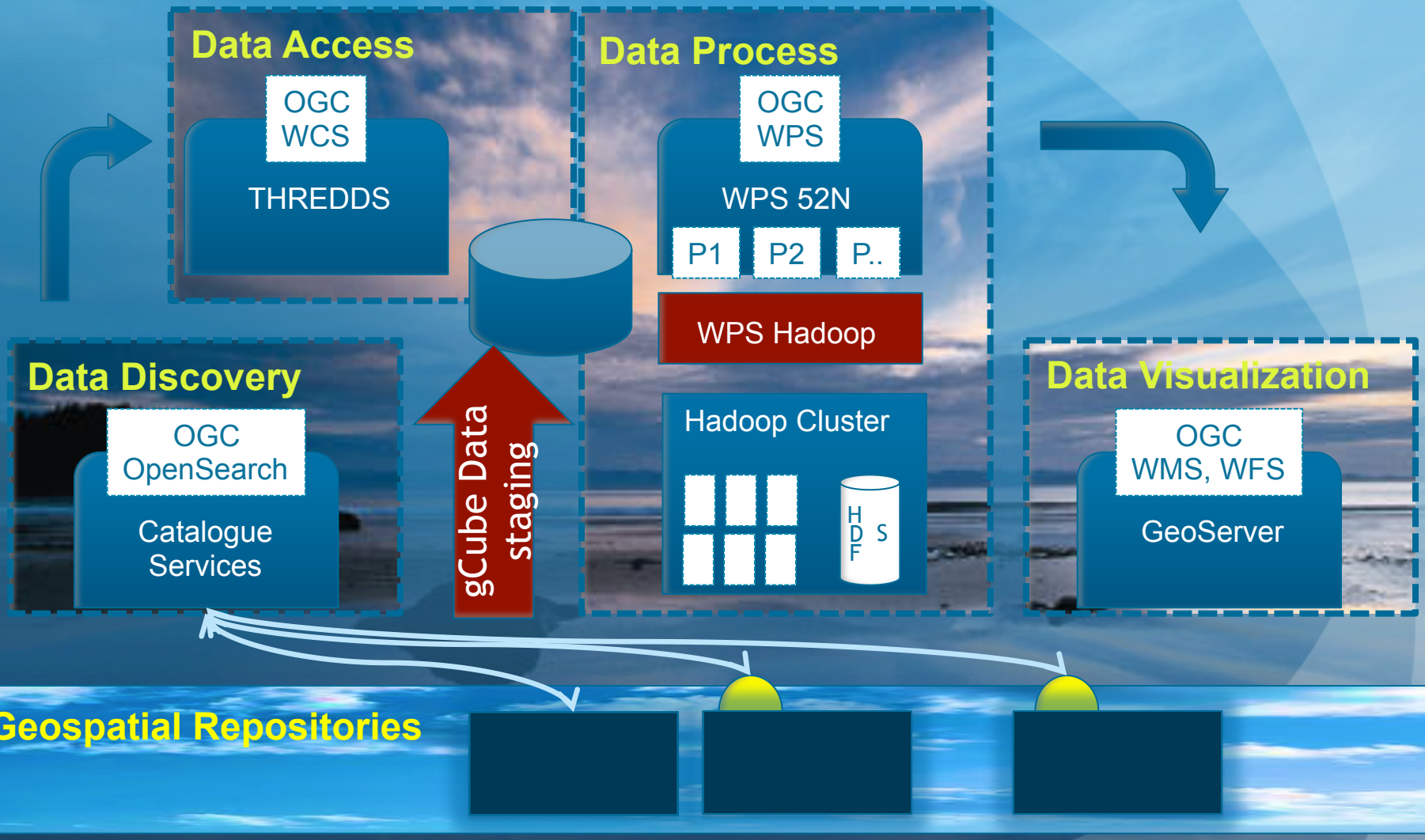
Access
- Web Coverage Service (OGC WCS)
- THREDDS: implements access protocols to netCDF (v. 4.2.20) data, OpenDAP (v 2.2.2), WCS

Process
- Web Processing Service (OGC WPS)
- 52North (2.0 RC8) framework: spatial resampling, temporal aggregation as WPS processes

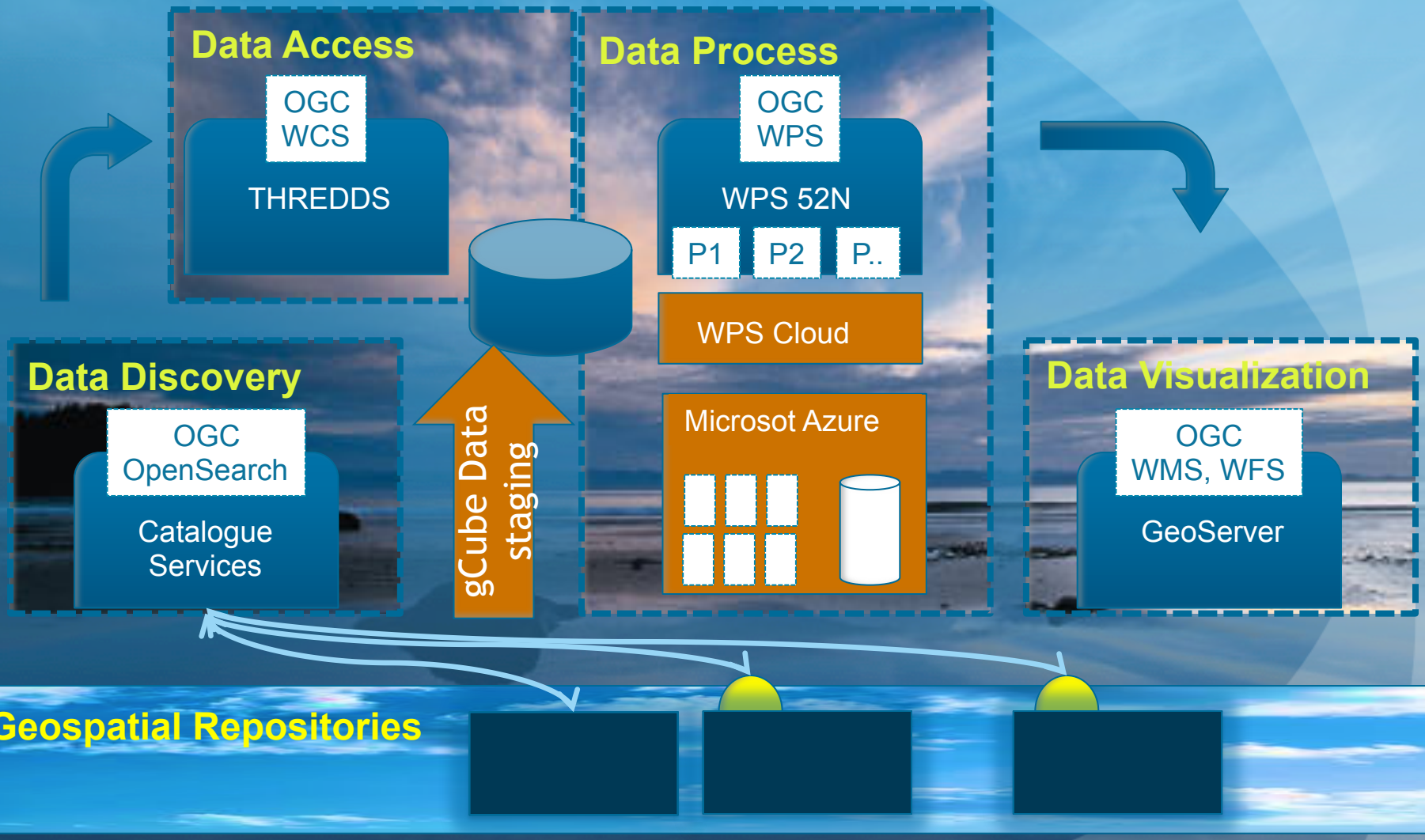
Computing
- Hadoop 0.20.2 (CDH3)
- WPS processes as map/reduce pure implementations

Publish and Visualize
- Web Map Service and Web Feature Service (OGC WMS, WFS)
- Geoserver, GeoTools (v. 2.7.4)

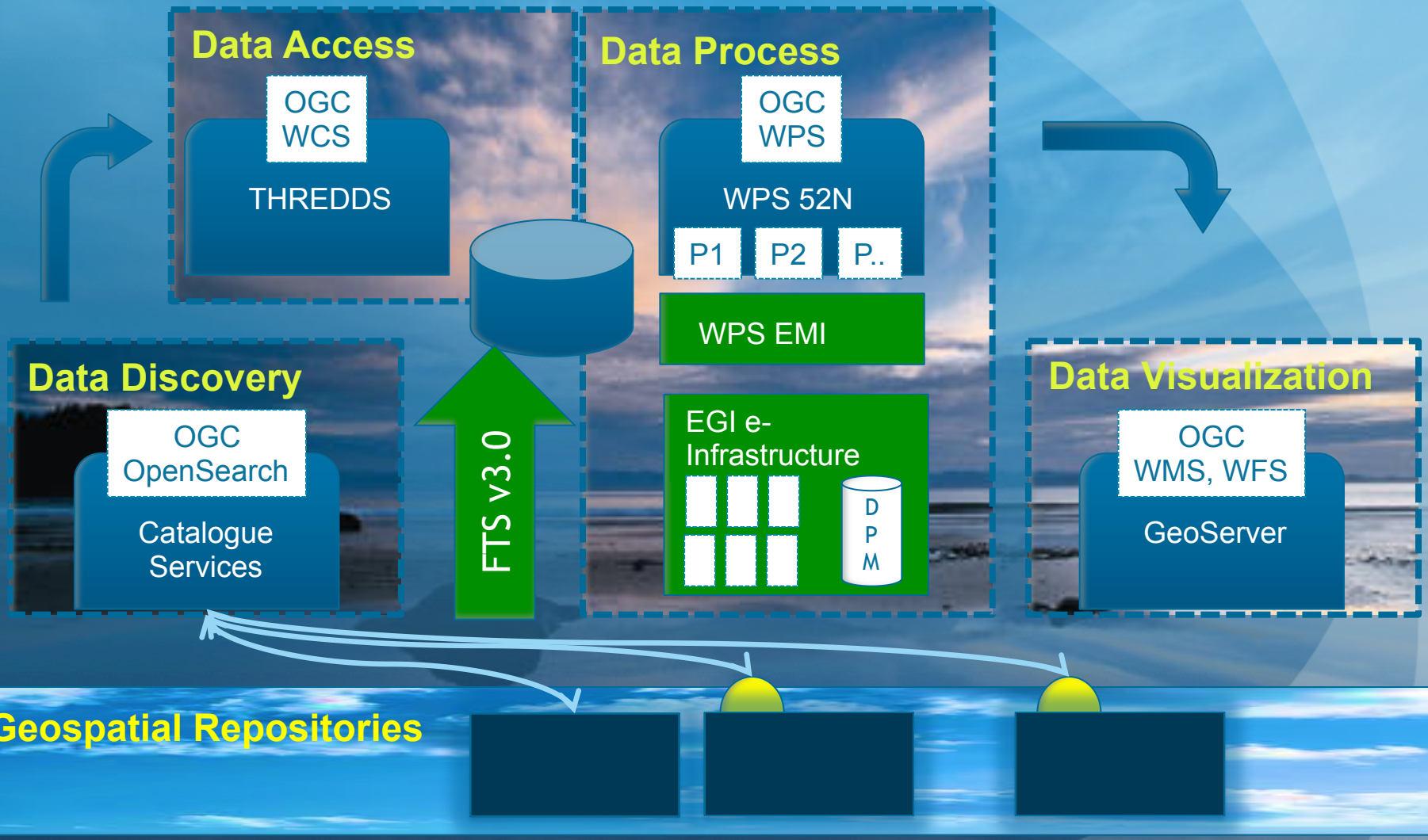
ENVRI and D4Science



ENVRI and EGI



ENVRI and EGI

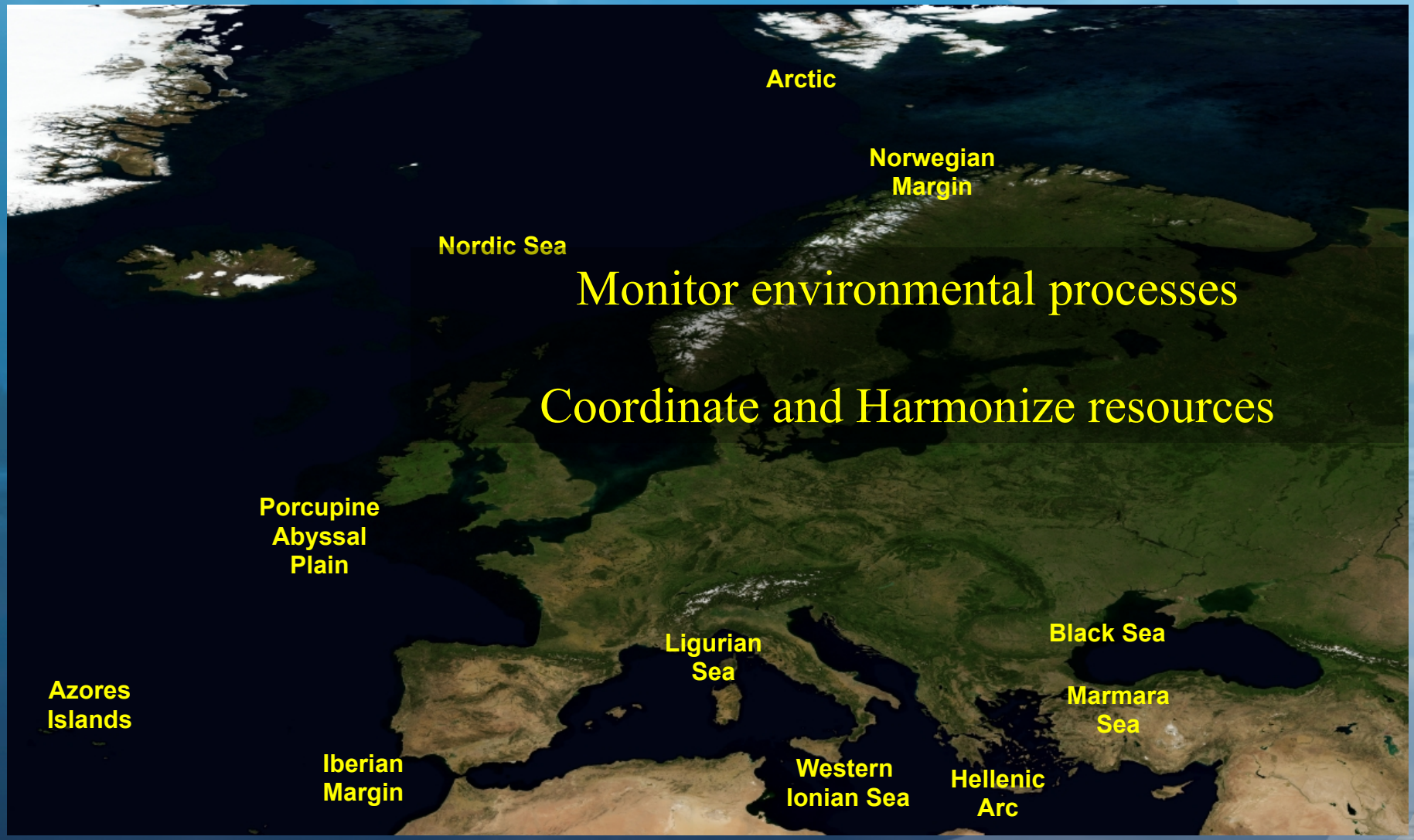


Follow us at

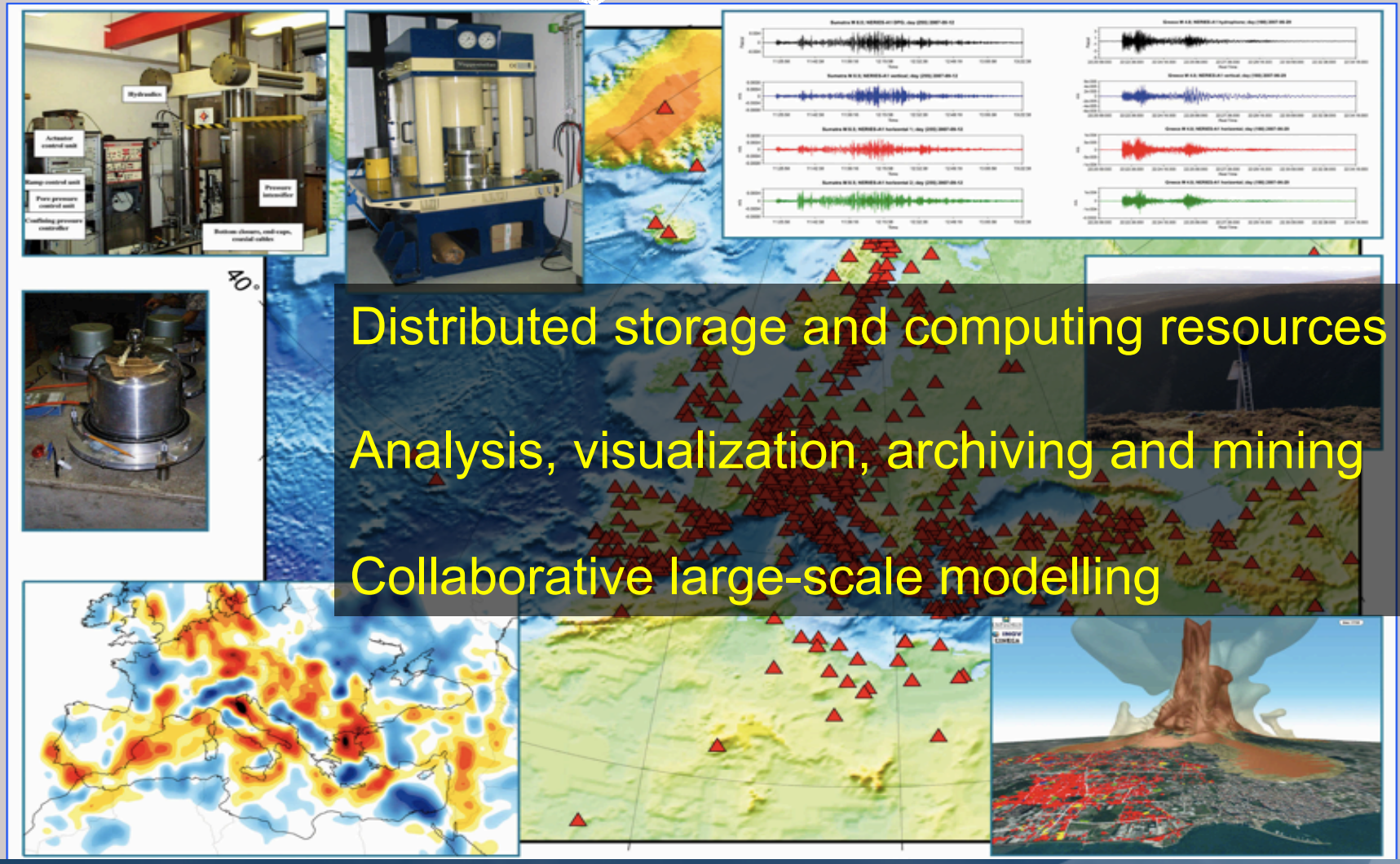
www.d4science.org
www.gcube-system.org
www.envri.eu

THANK YOU

EMSO: European network of underwater observatories



EPOS: Seismic and geodetic permanent national monitoring networks



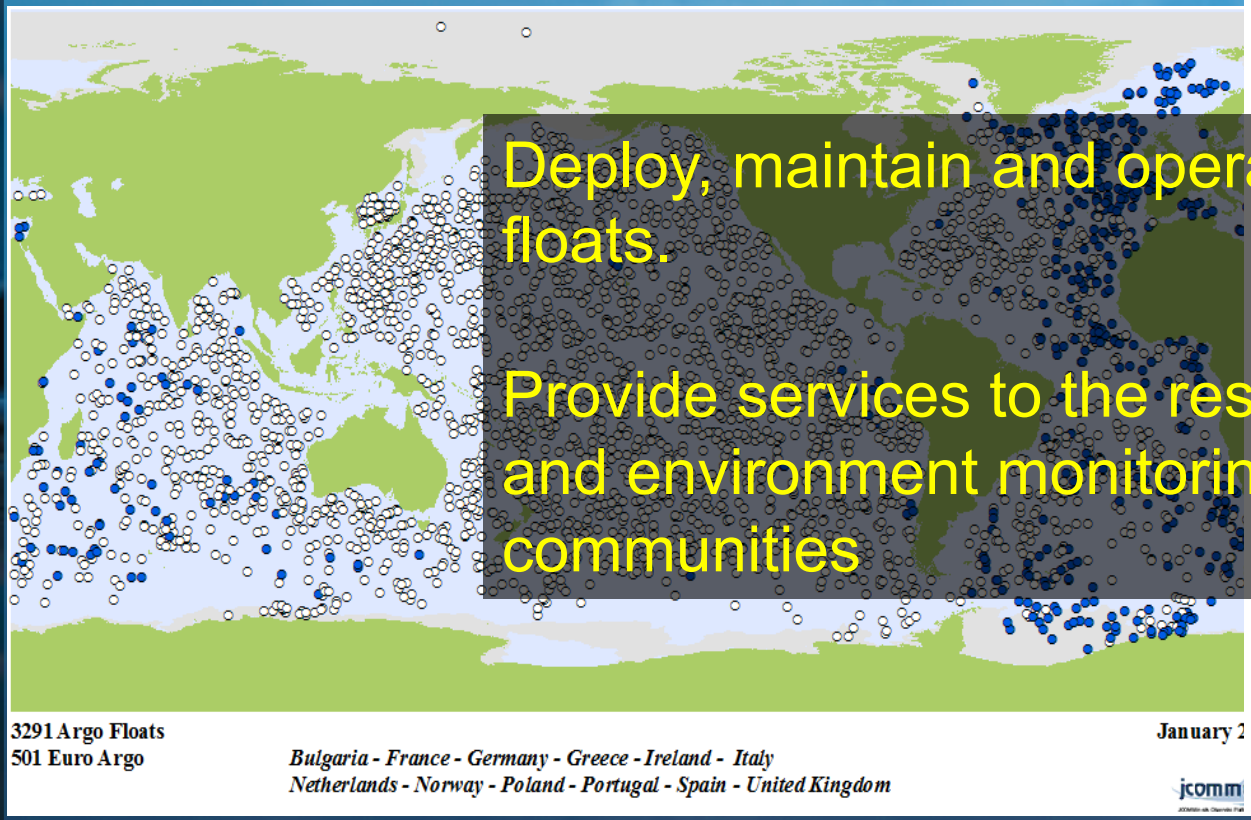
Distributed storage and computing resources

Analysis, visualization, archiving and mining

Collaborative large-scale modelling

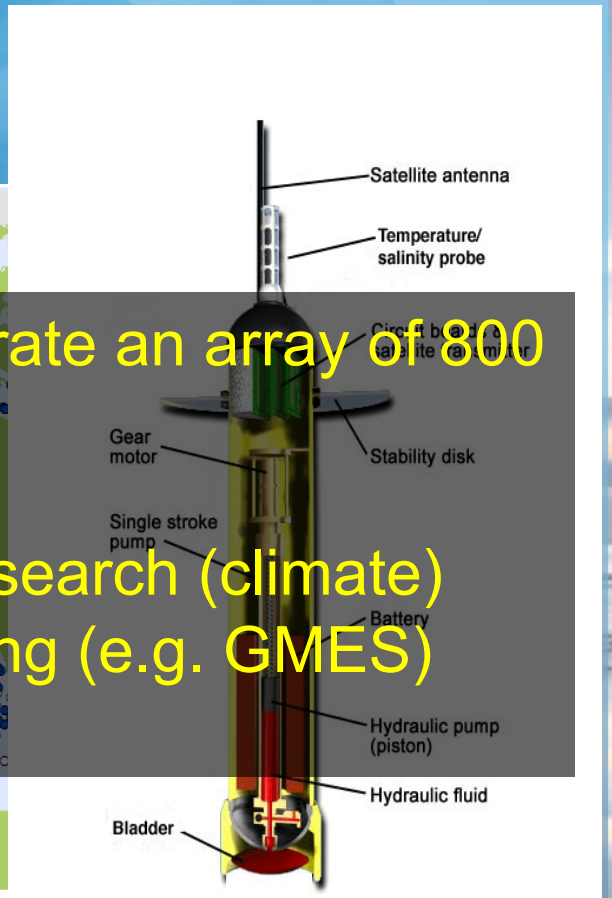
EURO ARGO: European component of a world wide in situ global ocean observing system

A dual use : research and environmental monitoring



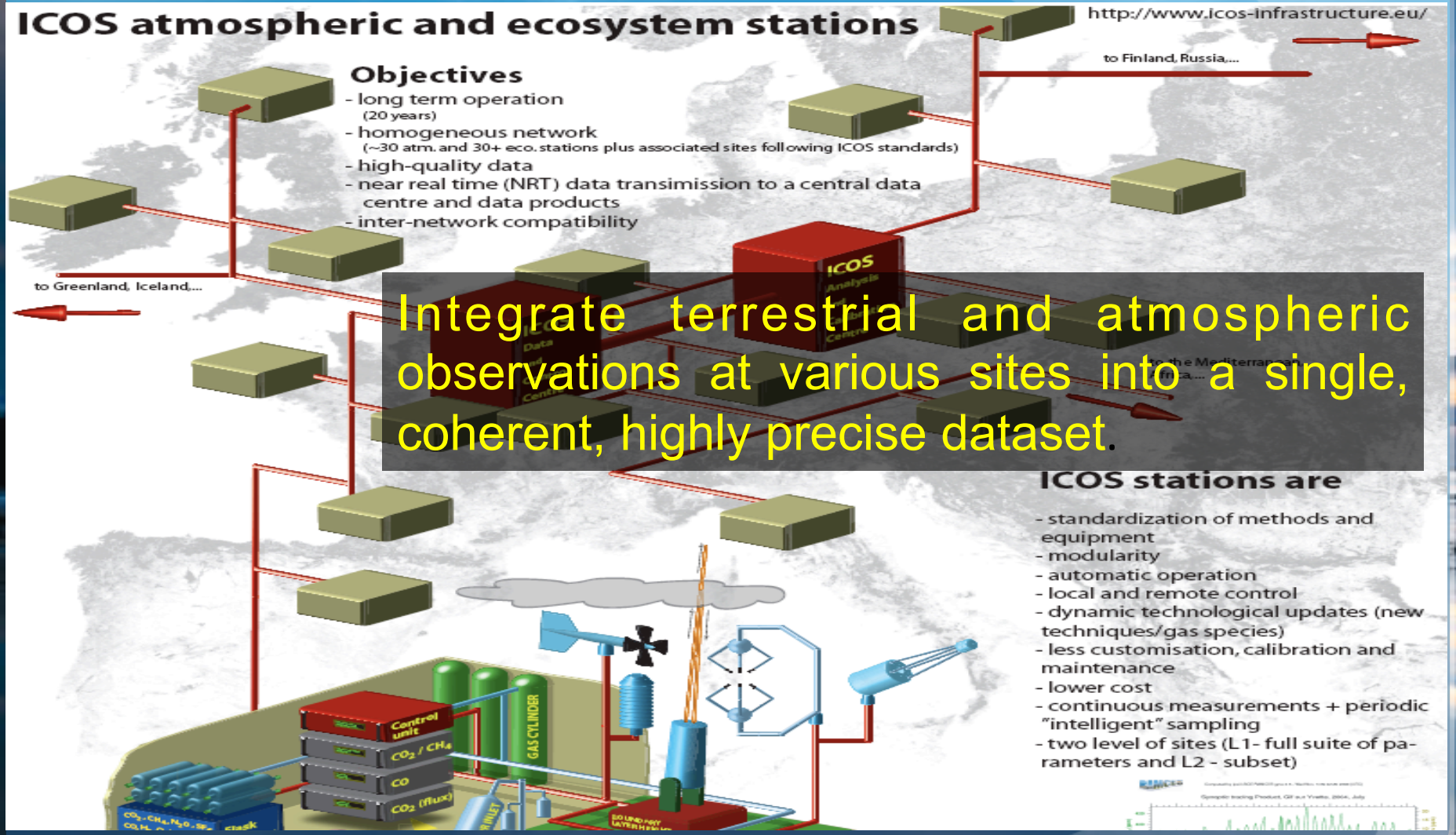
Deploy, maintain and operate an array of 800 floats.

Provide services to the research (climate) and environment monitoring (e.g. GMES) communities

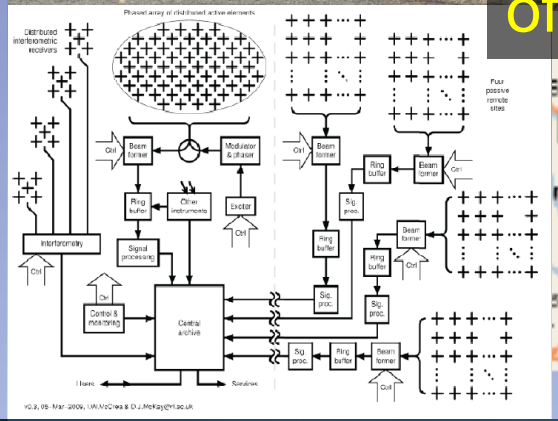


ICOS

Network of standardized high precision integrated stations

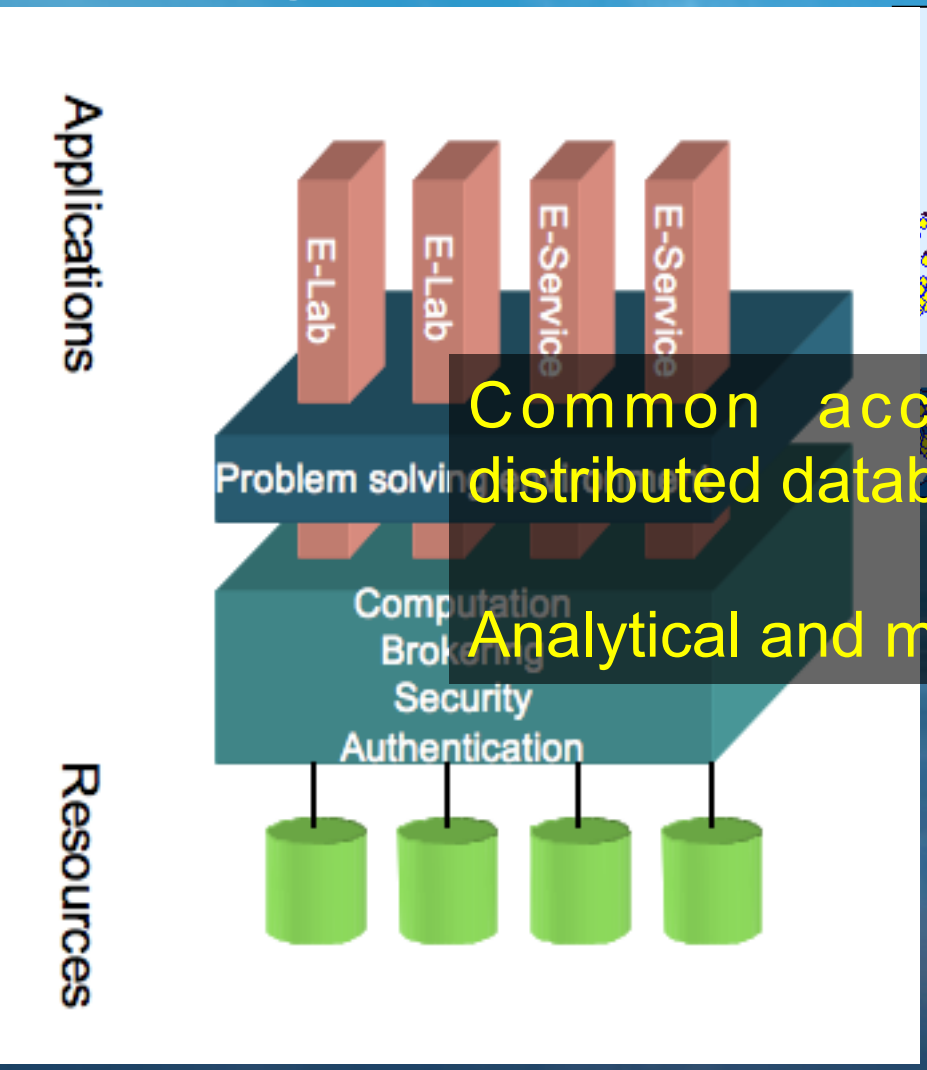


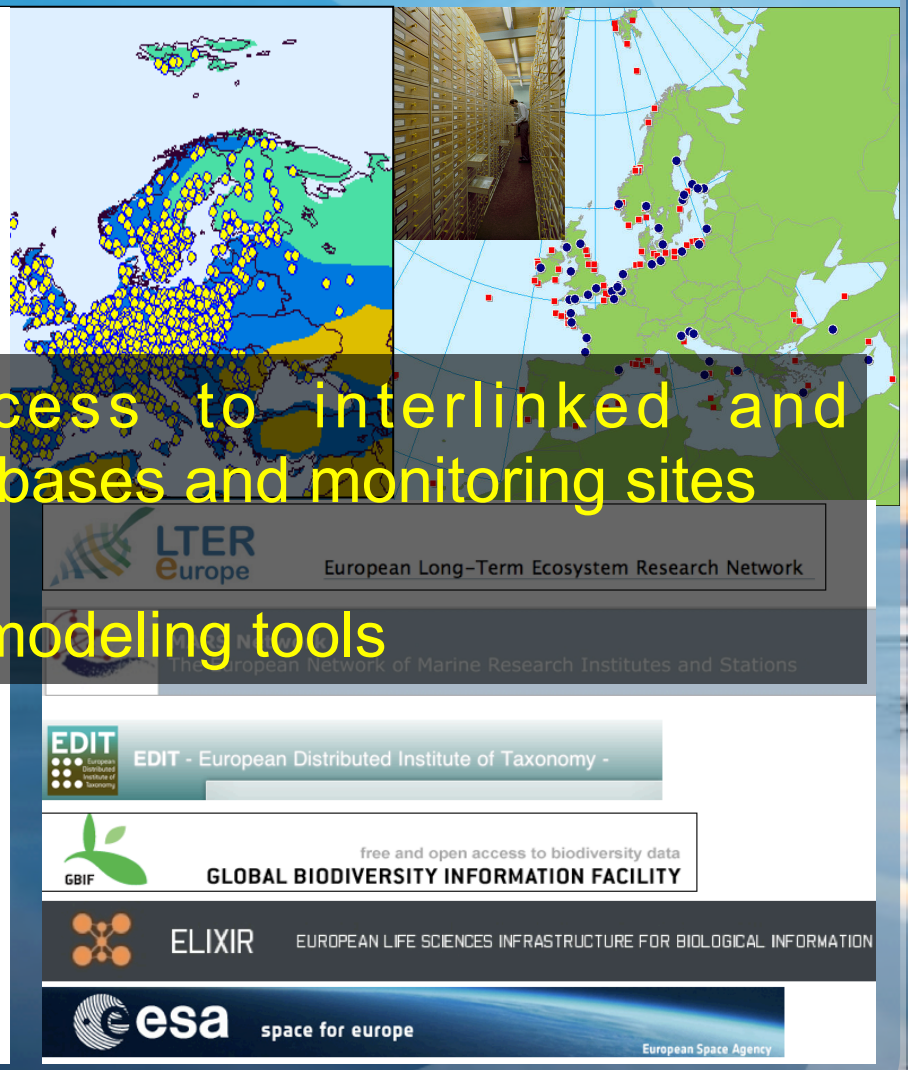
EISCAT_3D: System Design



Preparatory Phase to reach a sufficient level of maturity with respect to technical, legal and financial issues so that the construction of the EISCAT_3D radar system can begin

LIFEWATCH: European research infrastructure federating marine, terrestrial and freshwater observatories





Common access to interlinked and distributed databases and monitoring sites

Analytical and modeling tools