Big Data & Open data
Brussels, May 7-8 2014

EPOS PP
European Plate Observing System | FP7 Preparatory Phase Project

Big data and Open data for solid Earth science

Massimo Cocco & EPOS Team

Istituto Nazionale di Geofisica e Vulcanologia

EPOS PP Coordinator

INGV

www.epos-eu.org

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

Solid Earth Science

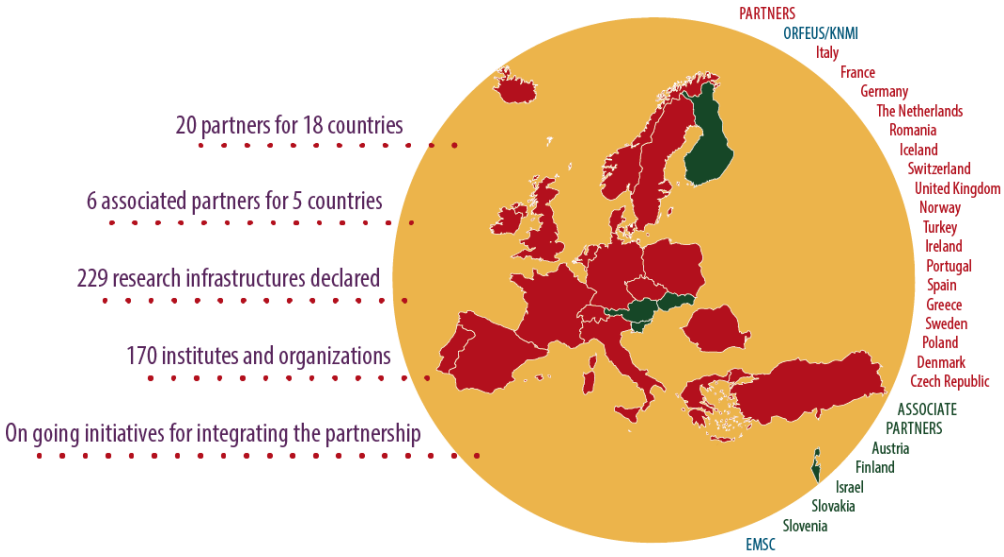
KEYWORDS

- **Multidisciplinary contributions**
- **Services to society**
- **Community building**
- **Geo-Hazards**
- **Geo-Resources**
- **Environmental changes**

European Plate Observing System: Mission

EPOS is a **long-term plan for the integration** of research infrastructures for solid Earth Science in Europe

EPOS will integrate the **existing** advanced European facilities into **a single, sustainable, distributed infrastructure** taking full advantage of new **e-science opportunities**

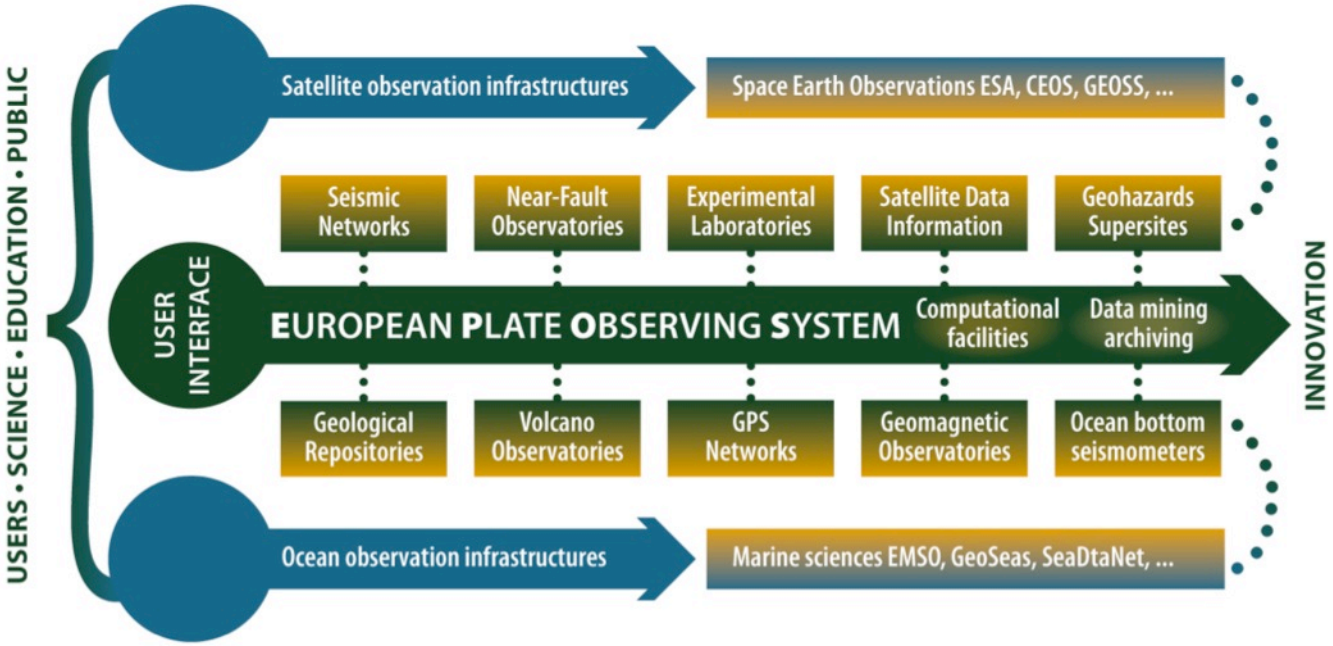


EPOS has the ambitious goal to facilitate research by providing **open access** to data, modeling tools, and facilities trough an efficient and multidisciplinary research platform.

This platform will facilitate innovative research for accurate, durable, and sustainable answers to societal questions relevant to the environment and human welfare.

EPOS Community

EPOS integrates a large number of infrastructures and communities



23 countries involved: **20** partners,
6 associated partners,
10 Working Groups (>350 members)

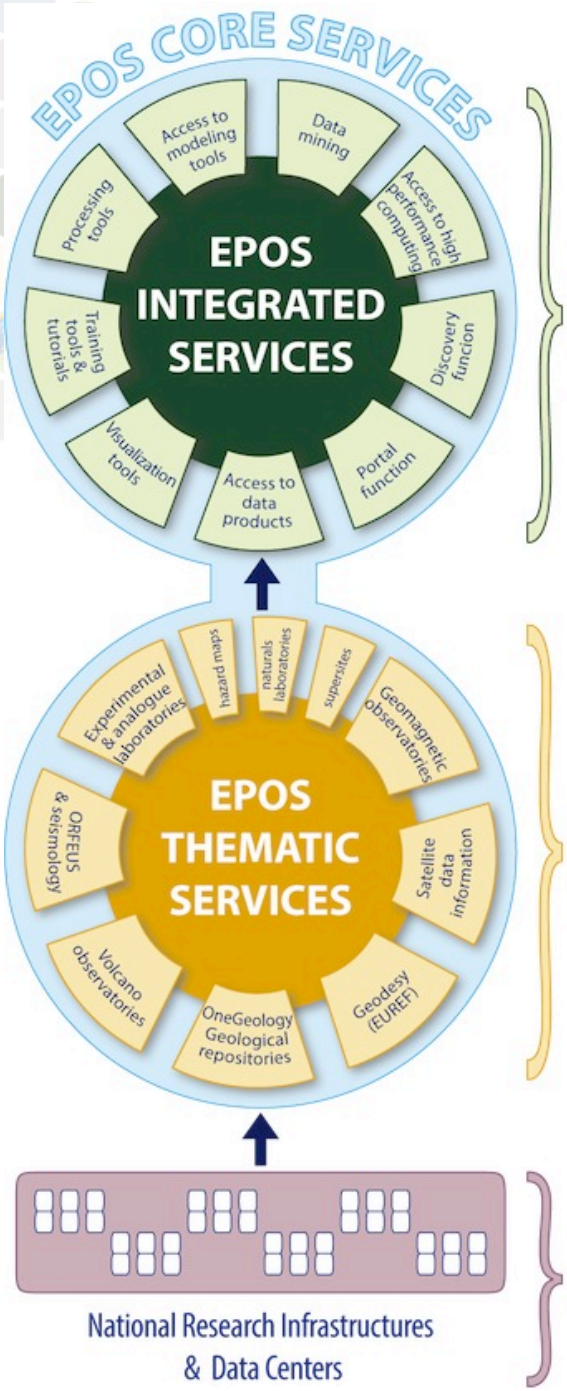
Research Infrastructures involved in EPOS
RIDE database: http://www.epos-eu.org/ride/

244 RIs - value (>350M€)
annual running costs (84M€ / per year)

138 Institutions
2272 GPS receivers | **4939** seismic stations
464 TB Seismic data | **1,1** PB Storage capacity
828 instruments in **118** Laboratories

EPOS will increase their efficiency, improve and simplify their use, and allow multilateral strategic coordination for their sustainability, operation, and development

Topological Architecture



The **EPOS Integrated Core Services** will provide access to multidisciplinary data, data products, synthetic data from simulations, processing and visualization tools, ...

The **EPOS Integrated Core Services** will serve scientists and other stakeholders, young researchers (training), professionals and industry

EPOS is more than a mere data portal: it will provide not just data but means to **integrate**, **analyze**, **compare**, **interpret** and **present** data and information about **Solid Earth**

Thematic Core Services are infrastructures to provide data services to specific communities (they can be international organizations, such as ORFEUS for seismology)

National Research Infrastructures and facilities provide services at national level and send data to the European thematic data infrastructures.

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

The National RIs

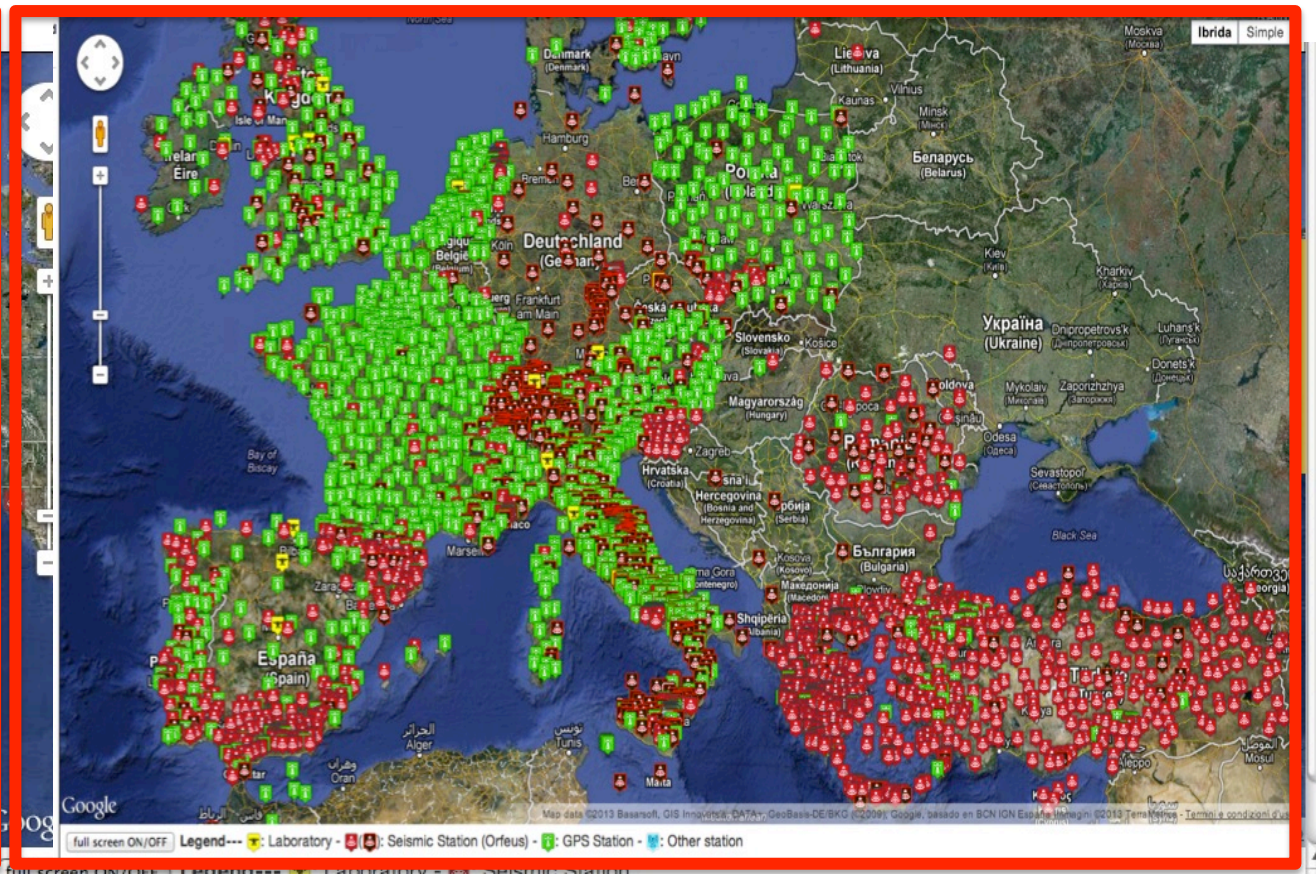
MAP OF:
- Seismic/GPS stations
- Laboratories
-- etc….

Diversity in data type and formats

http://www.epos-eu.org/ride/

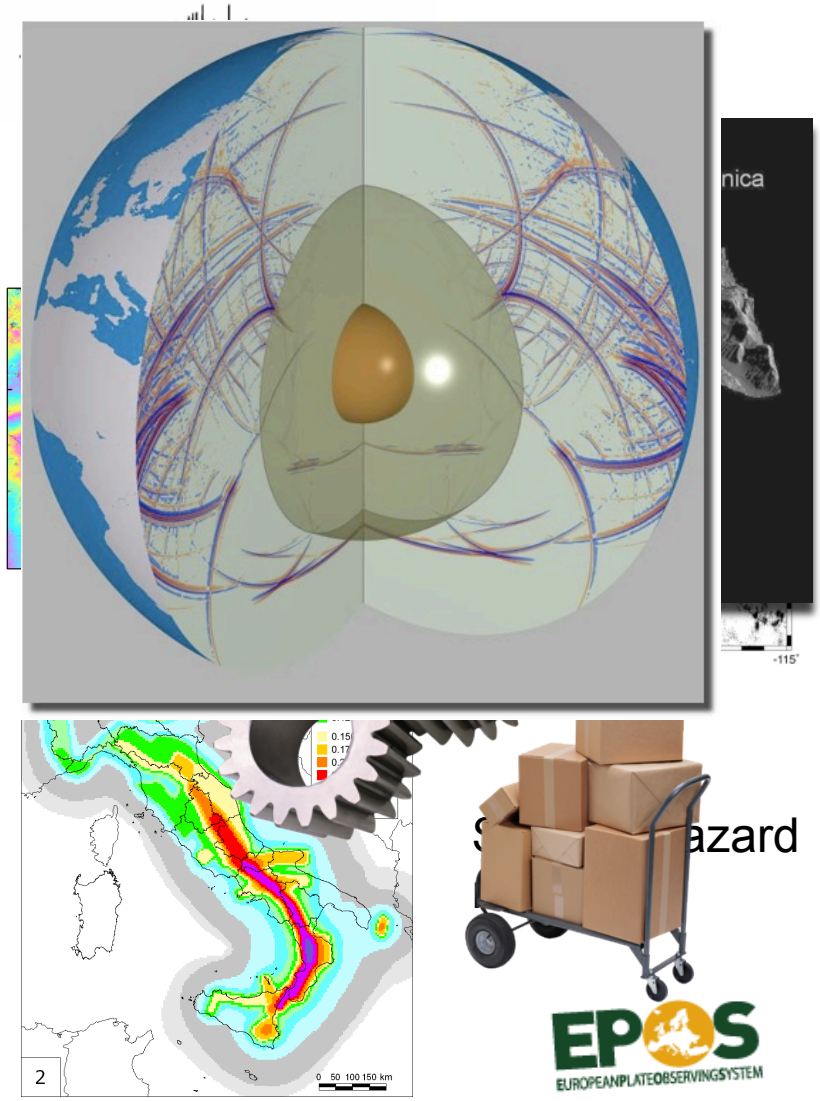
Research Infrastructure LIst

- 244 Research Infrastructures
- 138 Institutions
- 22 countries
- 2272 GPS receivers
- 4939 seismic stations
- 464 TB Seismic data
- 1.095 PB Storage capacity
- 828 instruments in 118 Laboratories



Access to Data Products (Taxonomy)

- **Level 0**: raw data, or basic data
- **Level 1**: data products coming from nearly automated procedures
- **Level 2**: data products resulting by scientists' investigations
- **Level 3**: integrated data products coming from complex analyses or community shared products
- **Level 4**. Software, IT tools



EPOS Data, Access, and IPR policy

Guiding principles: – open access
– licensing
– no charges

Protect EPOS legally | Unrestricted use & access | Trace EPOS use & users

Balance: Legal risk : Openness : Traceability

Licensing
IPR
Terms & Conditions
Restrictions

Data & Service **Providers** → Open Access deposit terms → EPOS → Open Access license → Data & Service **Users**

Categorization

as needed for legal aspects

Data & Data Products
Level 0,1, 2, 3

Tools & Software

Open : Restricted : Embargoed

mix and match as required

Anonymous : Registered : Authorized

Users

Big Data **Open Data**

Open Data & Services

- Accessibility (scientific use)
- Commercial use
- Dissemination to Society
- Service to society

Implications

- Metrics (use & re-use)
- Public funding
- Education & traning
- Ethic Issues

Functional Architecture

Higher level data products

INTEGRATED CORE SERVICES

API

EPOS PORTAL

METADATA CATALOGUE

SYSTEM MANAGER

SERVICES

DISTRIBUTED ICS

COMPATIBILITY LAYER

XML SEISMOLOGICAL SERVICES

GSAC XML SERVICES

SUPERSITES SERVICES

INDUCED S. CERIF XML

THEMATIC CORE SERVICES

SEISMIC DATACENTRES

GPS DATACENTRES

SUPERSITES E-INFRASTRUCTURES

INDUCED SEISMICITY DATACENTRES

National Networks

Seismic Networks

GPS Networks

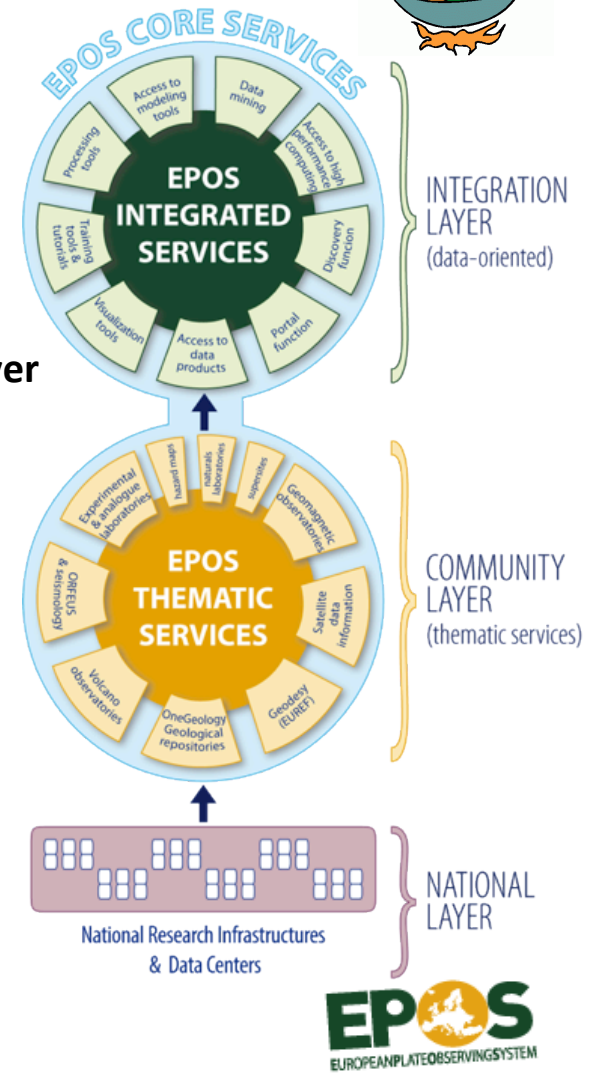
Geochemical Networks

Geomagnetic Data

Laboratories

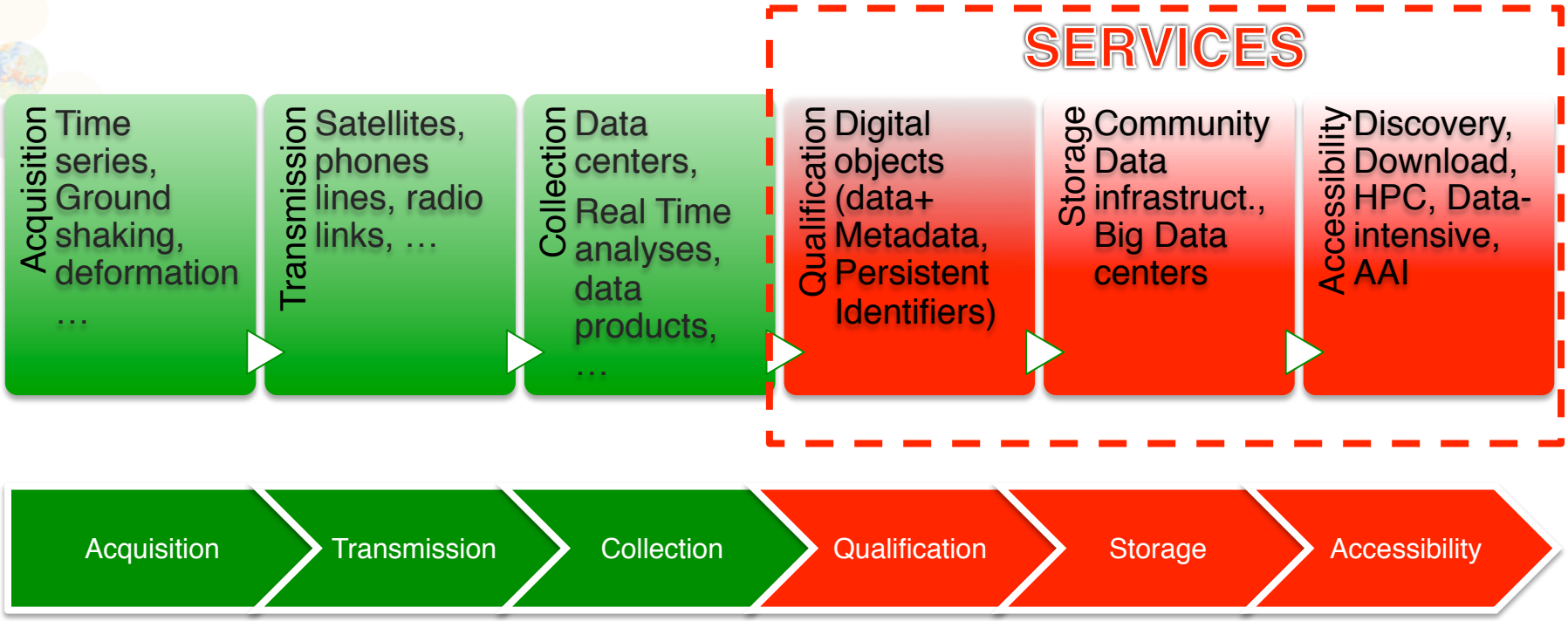
National Research Infrastructures

Compatibility Layer is the TCS-ICS Interface and it guarantees integration & interoperability



EPOS
EUROPEANPLATEOBSERVINGSYSTEM

Data Timeline

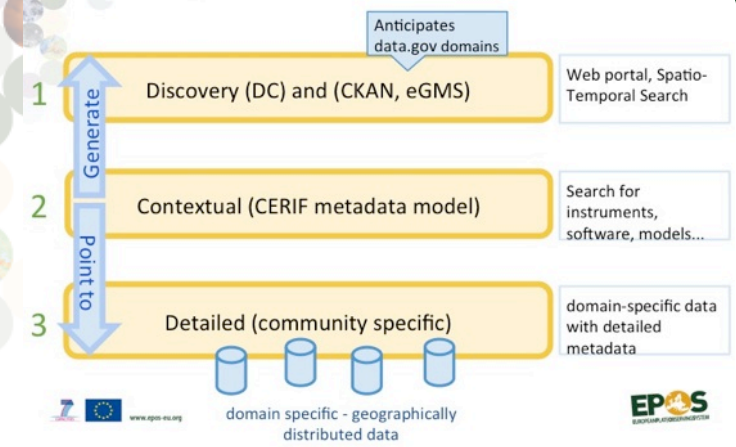


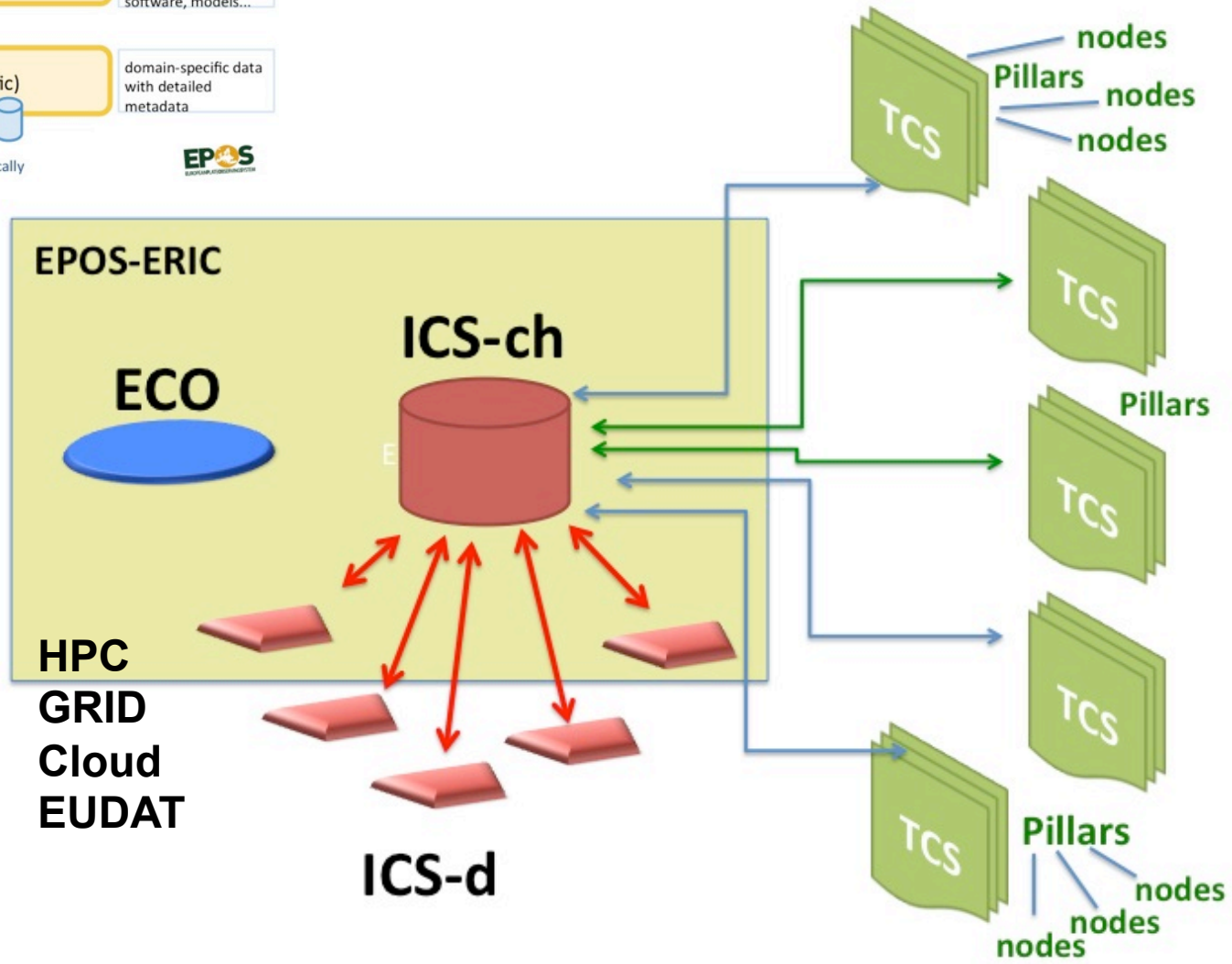
Data acquisition, validation & standardization

Data collection & preservation (PID, DOI)

Accessibility, integration, computation

EPOS Architecture





EPOS Challenges

- Providing services to solid Earth community
  - Engaging data providers & users (future data products providers)
- Involve other scientific communities
  - Environmental science (marine, atmosphere, ...)
- e-science community
  - IT innovation for developing e-RIs
  - Access to services for distributed resources (different timelines)
- Involve private sector with a clear strategy

A Paradigm Shift: from Data Driven to Data intensive Research

The earthquake data-driven research has entered a fundamental paradigm shift.

Data intensive applications.

To exploit the full potential of this rapidly growing European and Global data-rich environment,

To guarantee optimal operation and design of the high-cost monitoring facilities,.

Data-intensive research is rapidly spreading in the community.

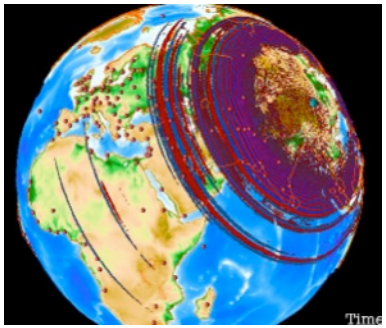
Large volumes of time-continuous seismograms contain a wealth of hidden information about the Earth's interior properties and wave sources, and their variation through time.

Mining, analyzing and modelling, this abundance of digital data will reveal new insights at all depths in the planetary interior and at higher resolution than is possible by any other approach.

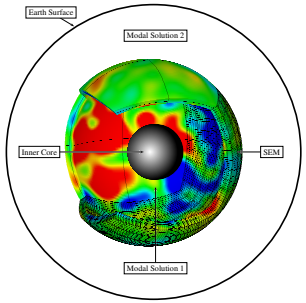
VERCE : Virtual Earthquake and Seismology Research Community e-science environment in Europe

Data Intensive simulation and inversion

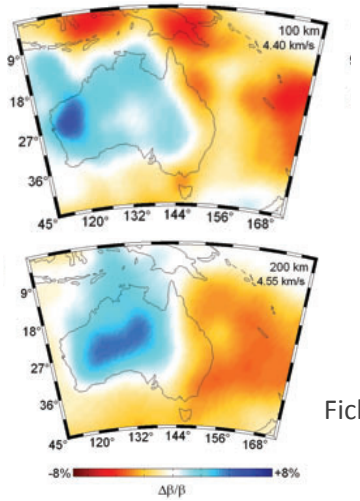
Seismic wave propagation and tomography



Komatisch *et al.* (2009)



Capdeville *et al.* (2003)



Fichtner *et al.* (2009)

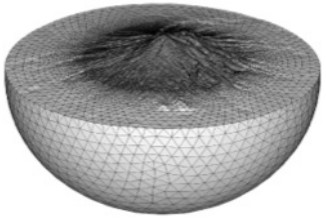
Global scale:

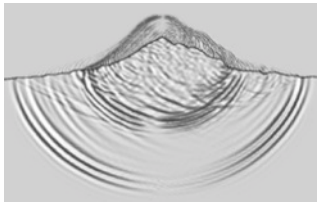
- **Waveform prediction for large earthquakes**
- **Full waveform inversion tomography: new inside in the deep Earth**

Regional scale:

- **Wave propagation in complex geological media**
- Full waveform inversion
- **Extended earthquake sources imaging**

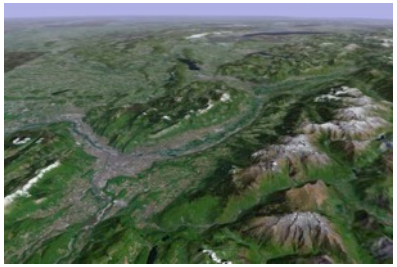
Aero-acoustic wave simulation in a volcano

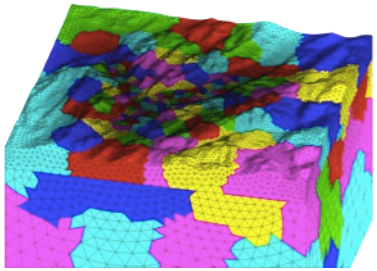


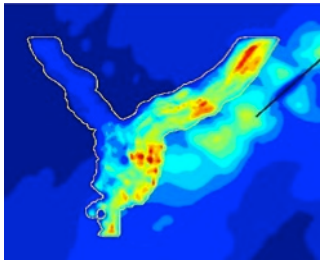


Käser *et al.* (2009)

Strong motion simulation: Grenoble Valley



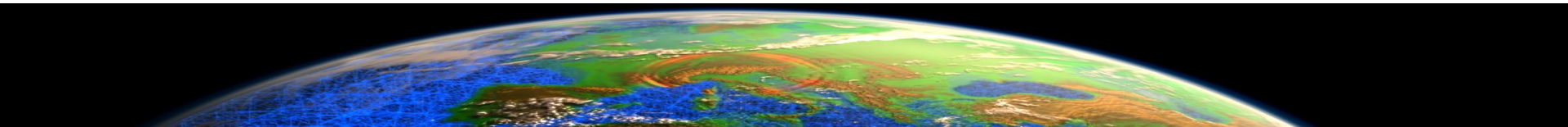




Chaljub *et al.* (2009); Delavaud *et al.* (2009), Käser *et al.* (2009)

Strong motion prediction:

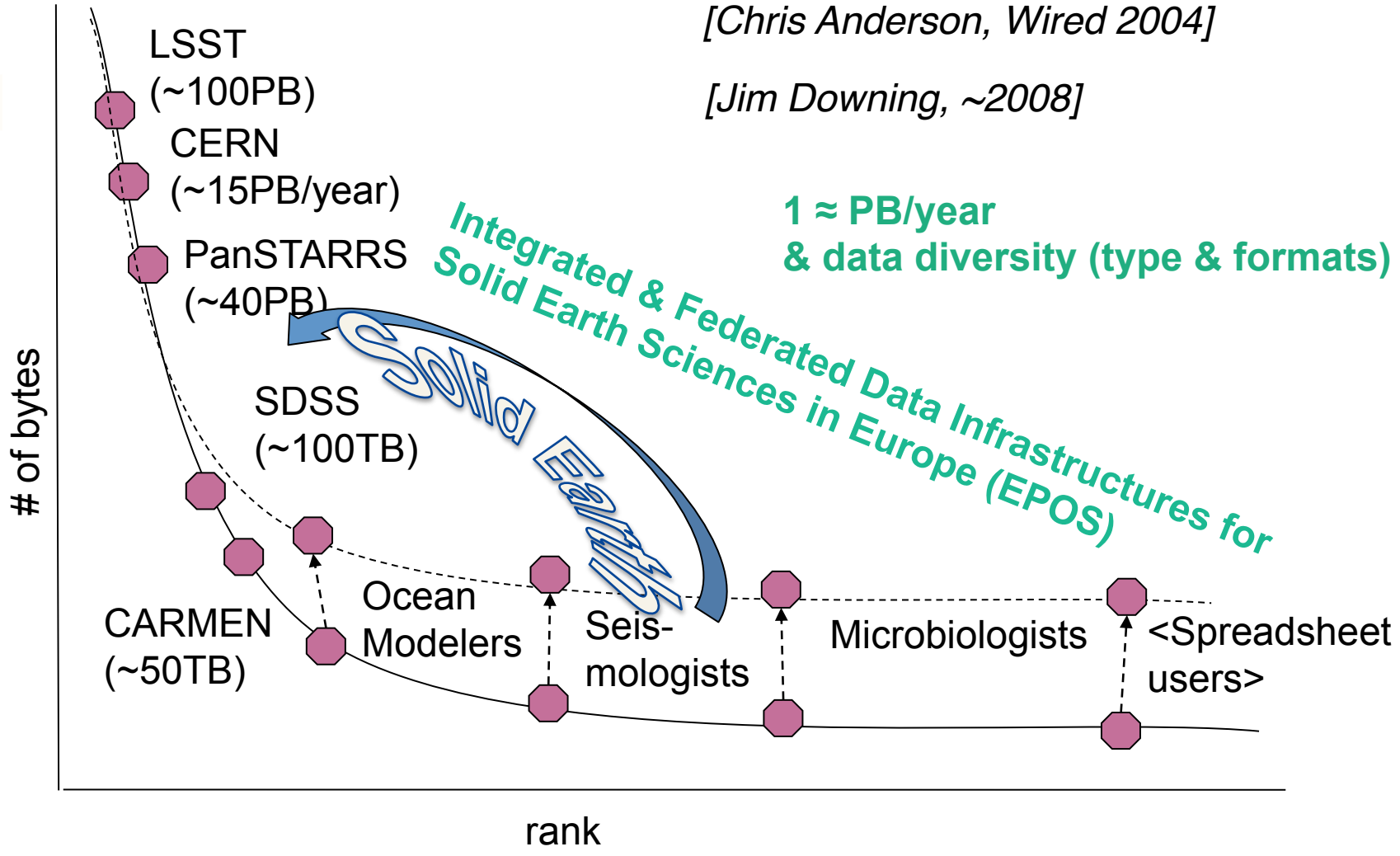
- **Physically-based hazard assessment**
- **Earthquake source dynamics**
- **Stochastic wave simulation**



The long "Heavy" tail

[Chris Anderson, Wired 2004]

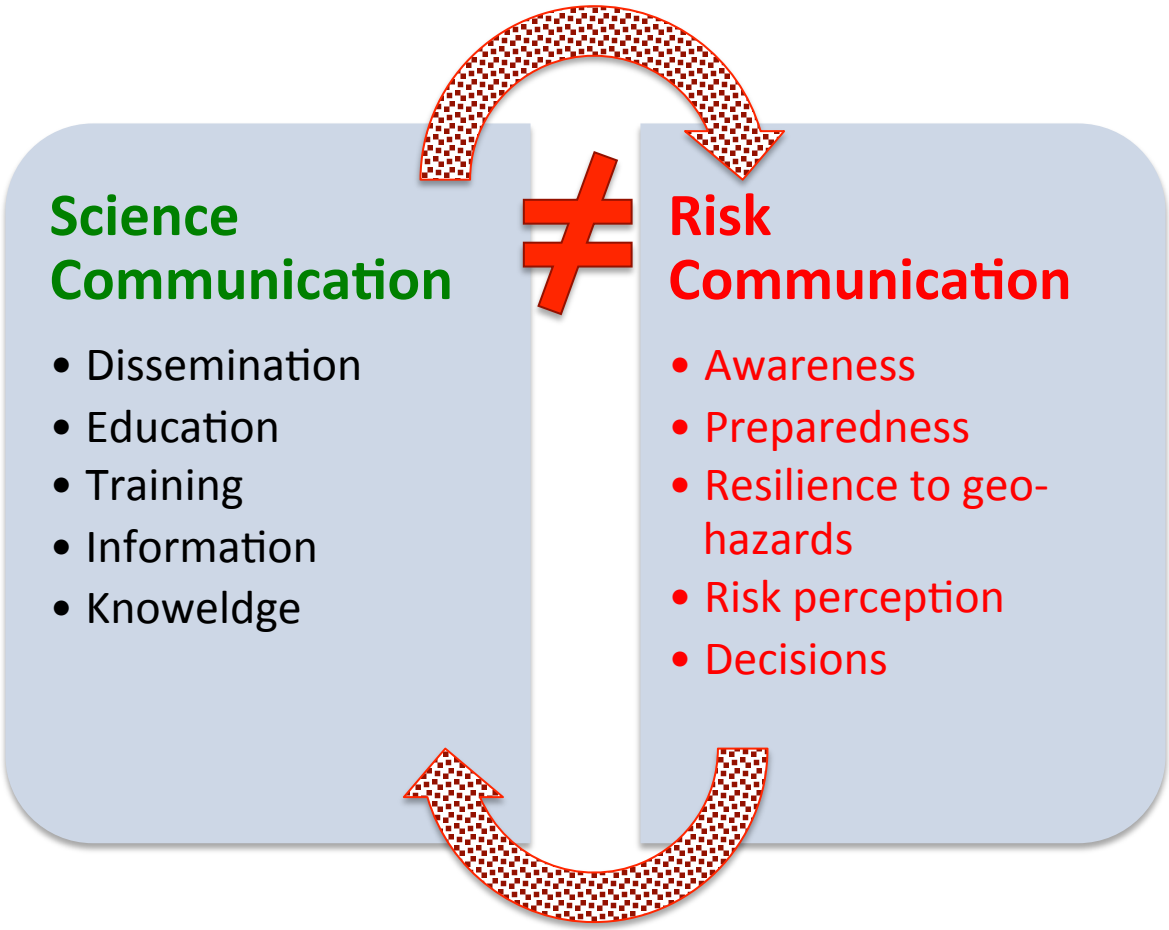
[Jim Downing, ~2008]



Integrated & Federated Data Infrastructures for Solid Earth Sciences in Europe (EPOS)

1 ≈ PB/year
& data diversity (type & formats)

Ethic Issues



Summary

- Individual **communities have their own thematic services** developed throughout many years and, in general, they are happy with them (!) ➜ ***ad hoc* solutions**
- In solid Earth sciences (EPOS), **data sharing has enormous potential** but there may not yet be enough consciousness of the scientific problems that can be addressed, i.e., a **new typology of scientists** targeting **multidisciplinary problems** is to be formed
- Building an e-infrastructure is very demanding given the diversification of the communities in terms of **different levels** of **data organization development/maturity** and **willingness** to be part of
- Must not loose pieces (communities) along the way ➜ **capitalize on** the **existing** developments and **introduce novelties** by making synergy with the different projects and the communities ➜ **efficient communication policy**.

Summary (cont'd)

- To achieve the best results, it needed **continuous orchestration** between scientific communities and ITs (e.g., scalability, AAI)
- EUDAT can represent a **data organization model and services** which can be instrumental toward EPOS e-infrastructures
- EUDAT and VERCE are posing particular attention to **large-to-huge data volumes analyses**
- The **communities** are undergoing a **positive**, **maturation process** and the **ITs** are **understanding progressively** the problems of the formers **and envisaging solutions ➜ mutual trust and synergy**
- **Interactions with industry** in Earth sciences require effective strategies and particular attention (**ethic issues**, use and re-use of scientific data)

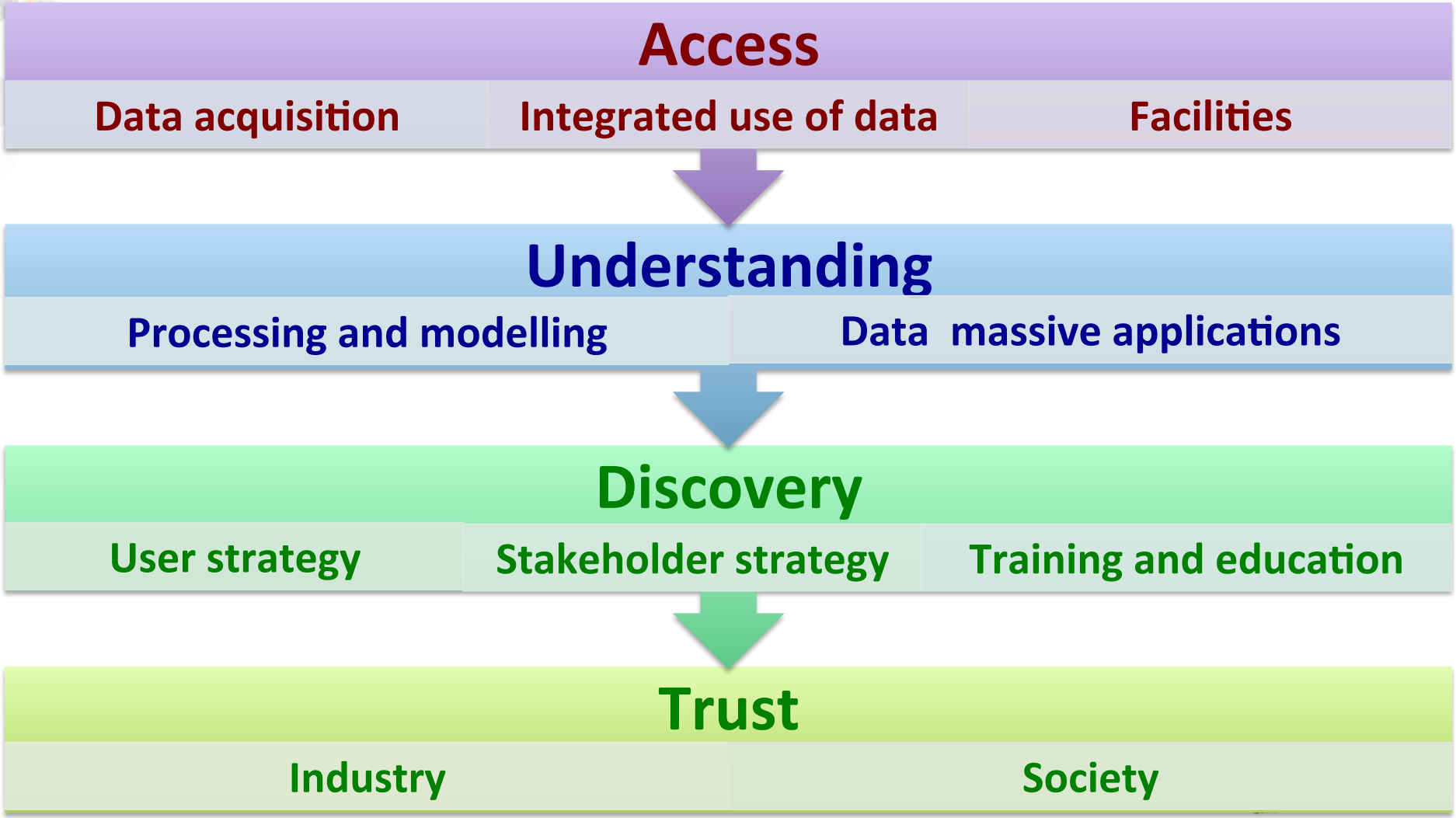
EPOS
EUROPEANPLATEOBSERVINGSYSTEM

Research infrastructures and e-science for data and observatories on Geo-Hazards and Geo-Resources

Thank you for attention

massimo.cocco@ingv.it www.epos-eu.org epos@ingv.it

The EPOS chain: high gain/high-*but manageable* risk



Comments on data sharing in EPOS

- EPOS (sub-)communities feature very **different levels** of **data organization development/maturity**
- Most communities have developed **in-house their own data services**
- Many communities are already **striving for their own data archive** and services and they are afraid and in some cases **difficult to share their data** *(e.g., why should I put resources in changing what I am doing if I can barely keep track of the services I am compelled to provide ?)*
- Many communities **think they have already the best services** *(i.e., they can carry out their own research!)* and they do not see why the data should be shared (or better qualified).
- Overall, it is a **slow process to introduce new concepts**, to adopt the **same jargon** and users/scientists often **not yet ready**
- **BUT it is a positive maturation process**

EPOS KEYWORDS

- **Integration** of the existing national and trans-national RIs
- **Interoperability** of thematic (community) services across several multidisciplinary communities
- **Open access** to a multidisciplinary research infrastructure for promoting cross-disciplinary research
- **Acknowledgment** of the data source
- **Progress in Science** through prompt and continuous availability of high quality data and the means to process and interpret them *(e.g., explore and mine large data volumes, results easily reproducible/replicable)*
- Data infrastructures and novel core services will contribute to **information**, **dissemination**, **education** and **training**.
- **Implementation** plans, which require strategic investment in research infrastructures at national and international levels.
- **Societal** contributions, e.g., hazard assessment and risk mitigation

EPOS Stakeholders

I. Data and service providers from the solid Earth sciences

(www.epos-eu.org/ride/)

- National data and service providers
- International data and service providers
- Data products providers

II. Scientific User Community

- Researchers from solid Earth Science
- Solid Earth science community projects (NERA, SHARE, REAKT, ...)
- Training and educational institutions, projects and initiatives
- Researchers and organizations from outside the solid Earth sciences

III. Governmental Organizations

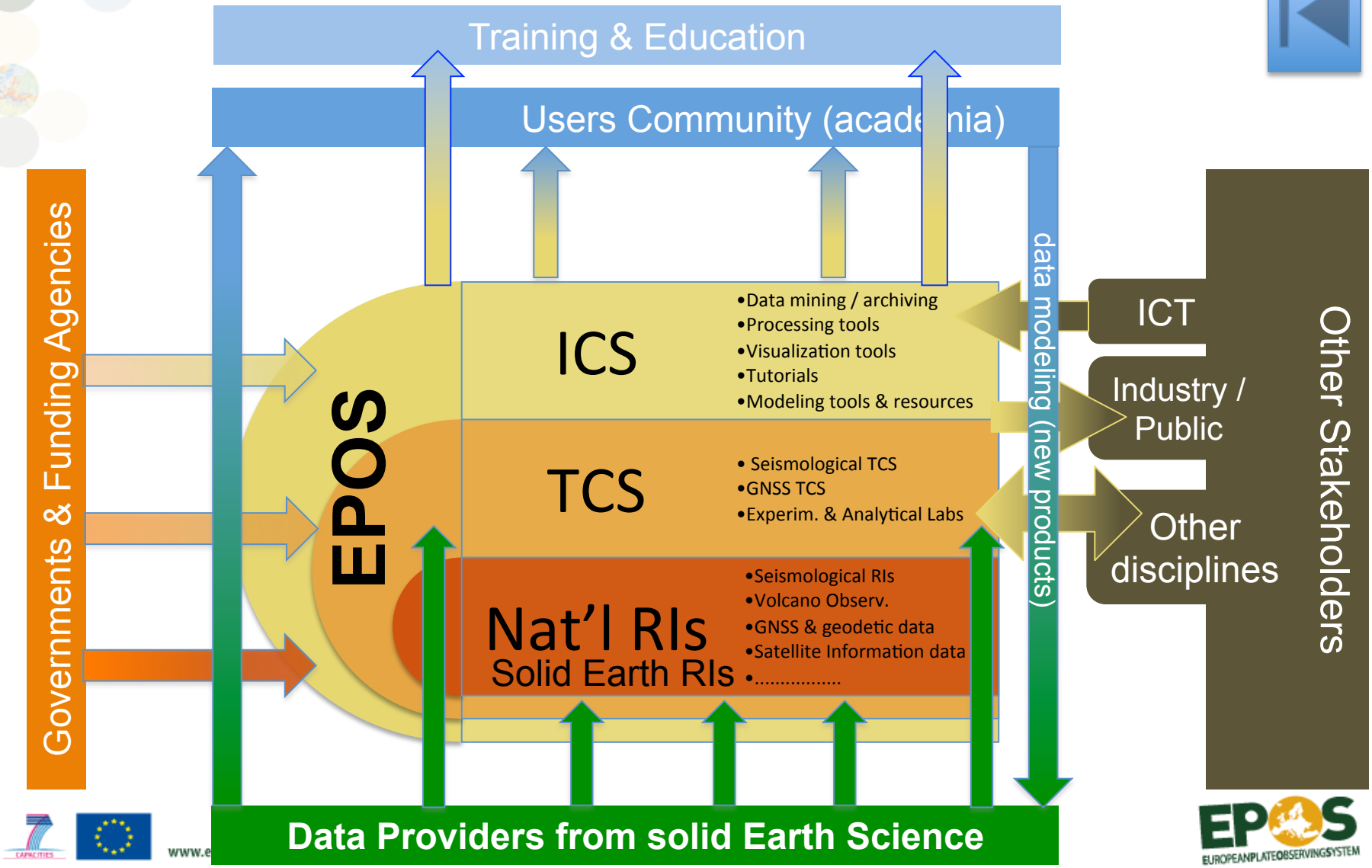
- National governments
- Funding agencies
- Civil protections authorities
- European Commission

IV. Other data and service providers and users

- IT projects and experts, Industry, Private data and service providers

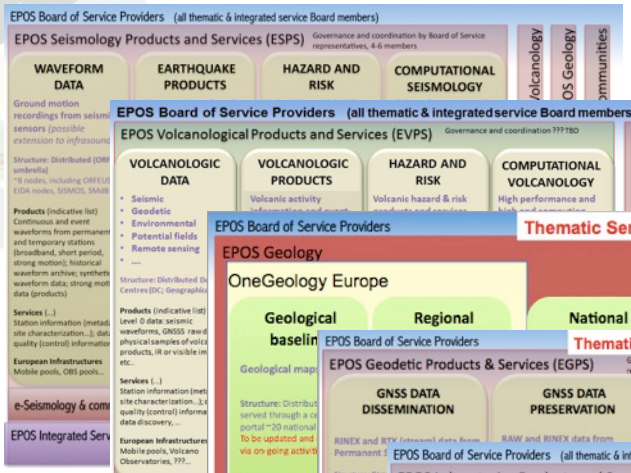
V. General Public

Stakeholder Strategy



Thematic Services (TCS)

- Governance
- Data Products
- Services



WG1 - Seismology

WG2 - Volcanology

WG3 – Geological Data

WG4 – GNSS Data

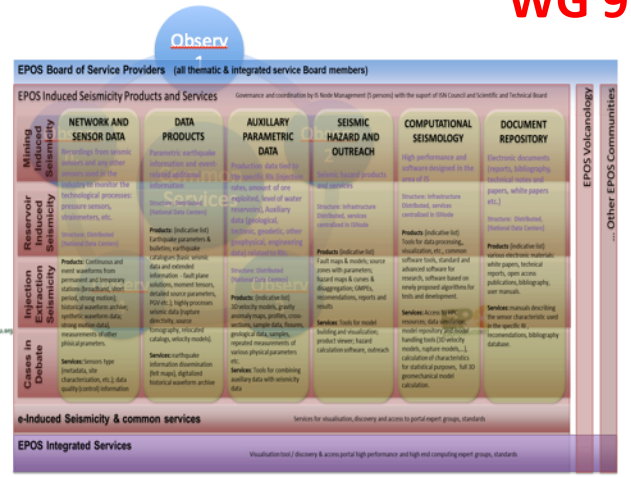
WG5
Near Fault
Observatories

WG6 – Analytical and Experimental Laboratories

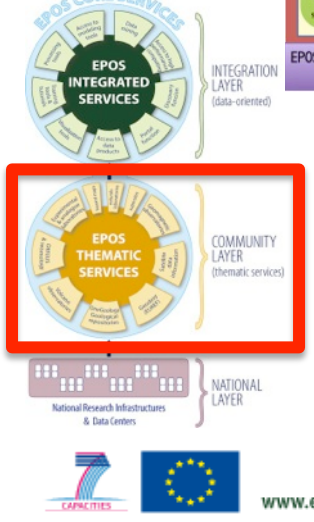
WG8 – Satellite Data

- WG9 a proxy of a network of distributed Ris

WG 9 – Geomagnetic Observ.



WG10 - Infrastructures for Georesources



www.epos-eu.org

EPOS EUROPEAN PLATE OBSERVING SYSTEM