Ways to improve Global Data Sharing and Re-use

Alberto Michelini

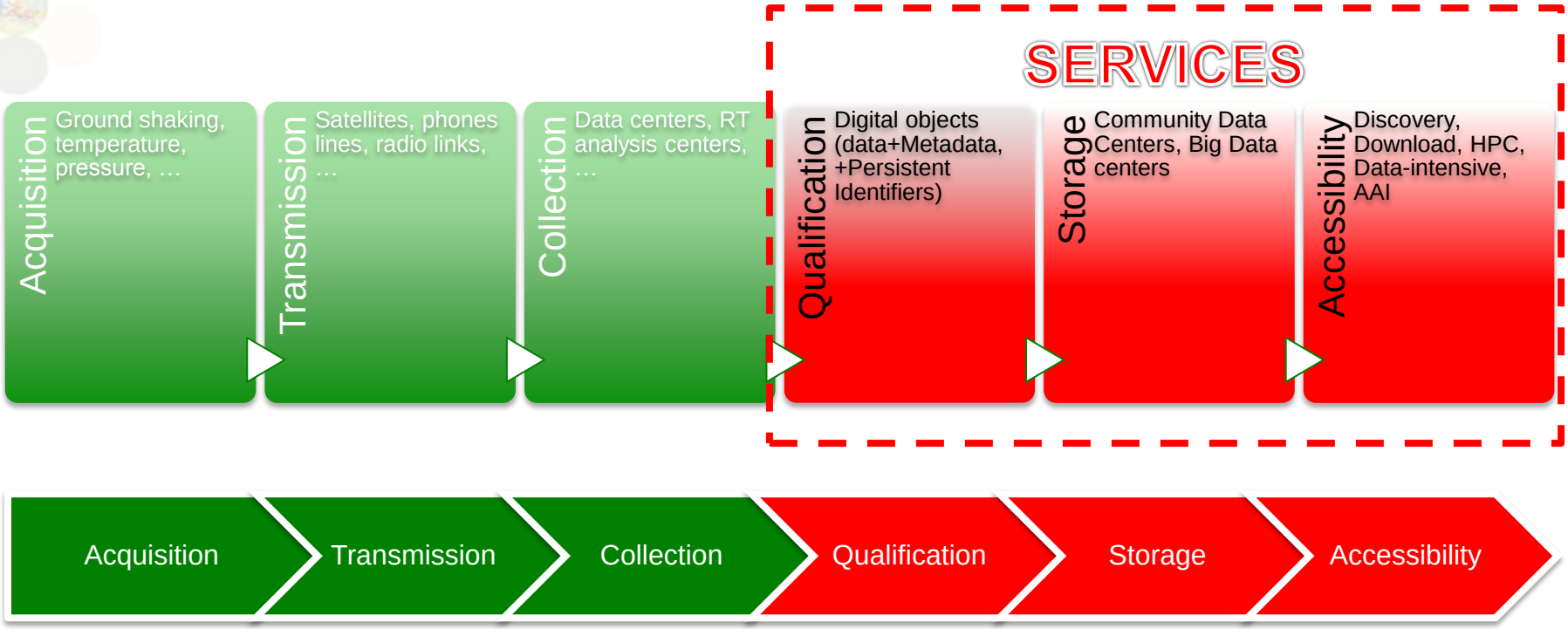
Istituto Nazionale di Geofisica e Vulcanologia

Rome, Italy

Who am I ?

- Seismologist (geophysicist)
  - Work with earthquakes at global scale (e.g., tsunami alert) where data sharing is routine/praxis
- Involved in EPOS (European Plate Observing System) - the ESFRI approved infrastructure for the solid Earth sciences
- Participating to the EC projects EUDAT and VERCE

Data Timeline



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

Talk structure

Global data and re-use perspective

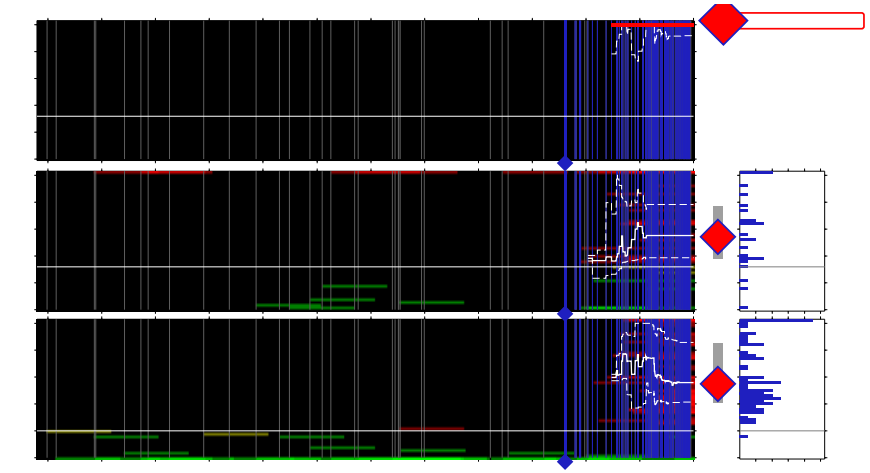
- From the side of a seismologist
- From the side of someone involved in EPOS
- From the side of someone involved in EUDAT

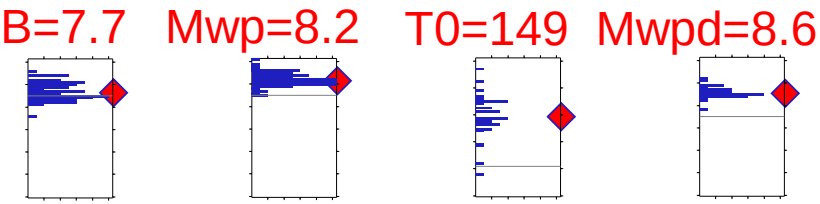
Lessons from recent Earthquakes

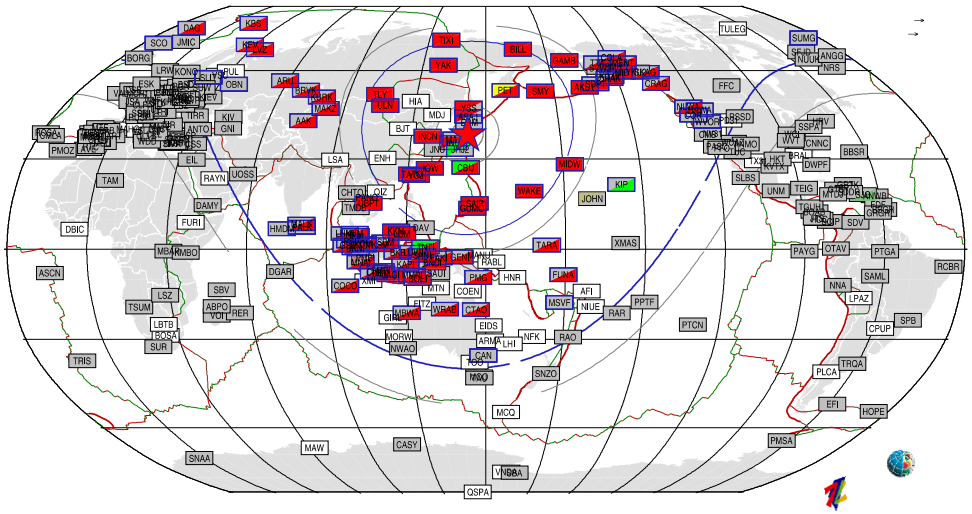
- Sumatra M 9.3 (Indonesia) 2004
- L'Aquila M 6.1 (Italy) 2009
- Haiti M 7.0 2010
- Maule M 8.8 (Chile) 2010
- Christchurch M 7.2 (New Zealand) 2010
- Tohoku M 9.0 (Japan) 2011
- Virginia M 5.8 (USA) 2011

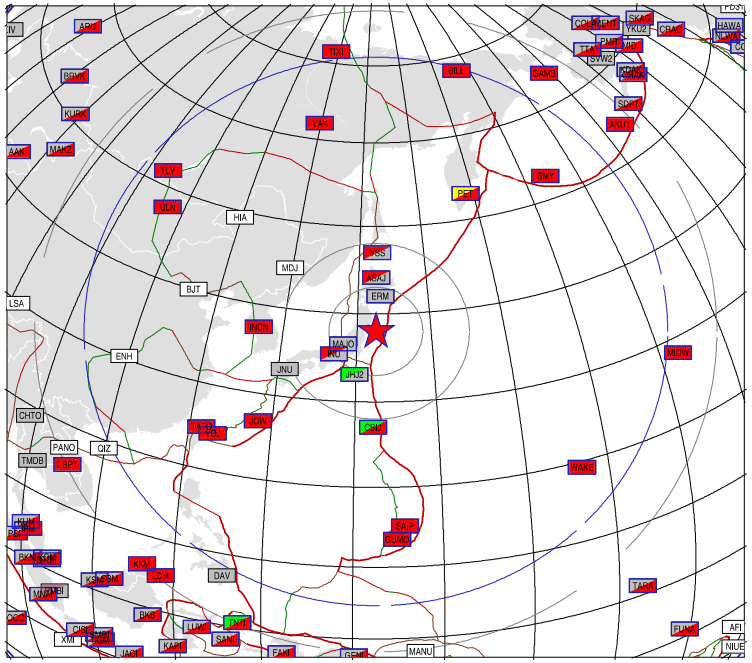
Tohoku Earthquake M9.0, 11 March 2011, 05:46:20 GMT (~12' from OT)

mB=7.7 Mwp=8.2 T0=149 Mwpd=8.6



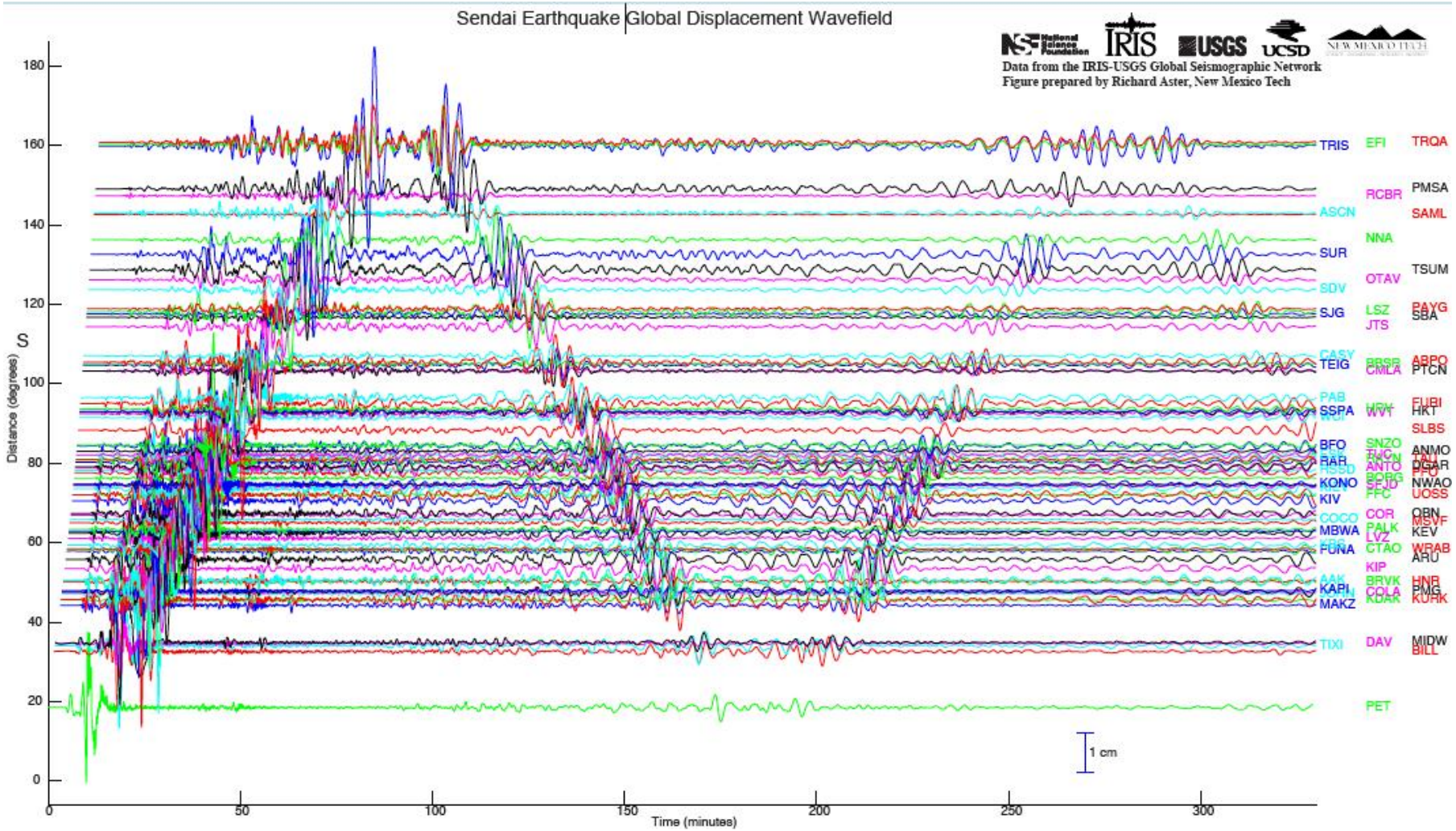






http://early-est.rm.ingv.it

Tohoku seismological global observations



Courtesy of Rick Aster, New Mexico Tech

Comments on data sharing in seismology

- ✓Global data ➜ **integration** (from national to global level integration for data and for services)
- ✓**Open access**
- ✓Real-time
  - Fundamental for seismic monitoring ➜ **societal impact** ➜ **information** & **dissemination**
- ✓Data organization accomplished with IT developments primarily WITHIN THE COMMUNITY (many years of investment) ➜ **interoperability**
- ✓**Progress in science** data promptly available ➜ rapid analysis ➜ improved earthquake knowledge available very shortly
  - Fantastic spin for **education** & **training**
- ✓Investments in research infrastructures (e.g., data, networks) and in the data centers ➜ **implementation**

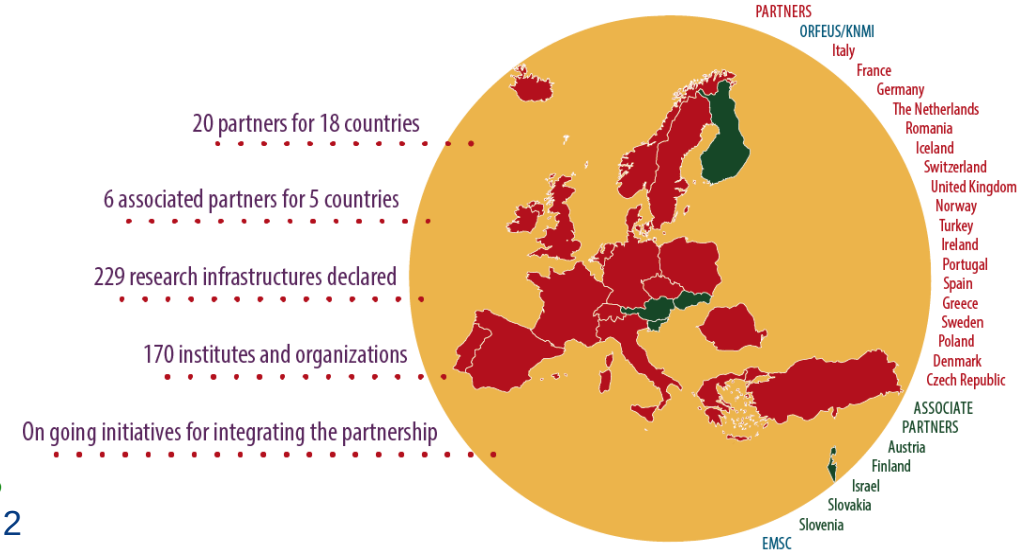
EPOS PP Mission

- The European Plate Observing System (**EPOS**) is a **long-term integrated research infrastructure plan** to promote innovative approaches for a better understanding of the physical processes controlling earthquakes, volcanic eruptions, unrest episodes and tsunamis as well as those driving tectonics and Earth surface dynamics
- EPOS aims at integrating the **existing advanced European facilities** into **one**, distributed **multidisciplinary Research Infrastructure** (RI) taking full advantage of new e-science opportunities
- The EPOS RI will allow geoscientists to study the causative processes acting from **10^{-3} s to 10^{6} years** and from ⌈ **m to 10^{3} km**

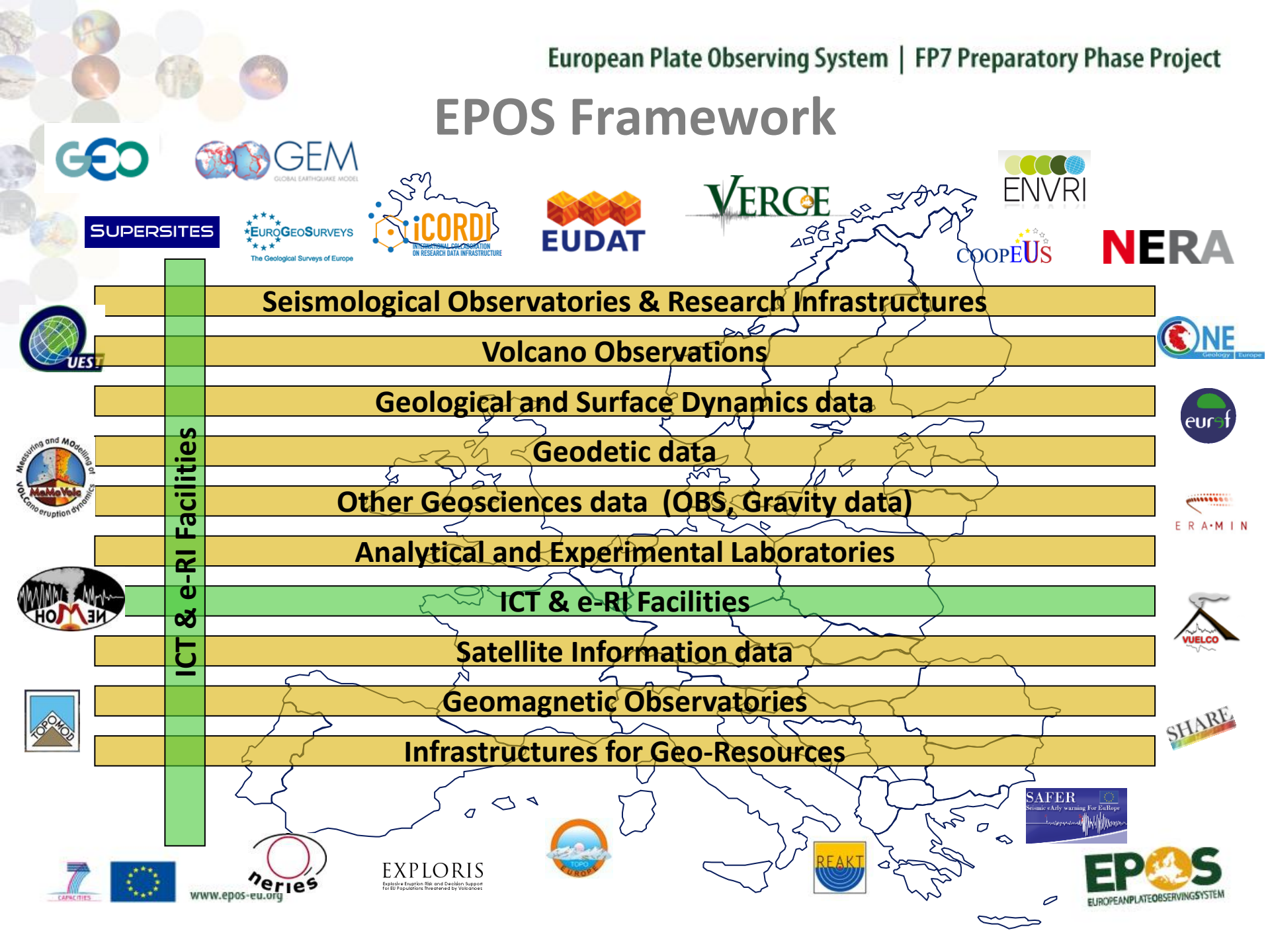
EPOS PP Timeline

www.epos-eu.org

mid-way through the PP



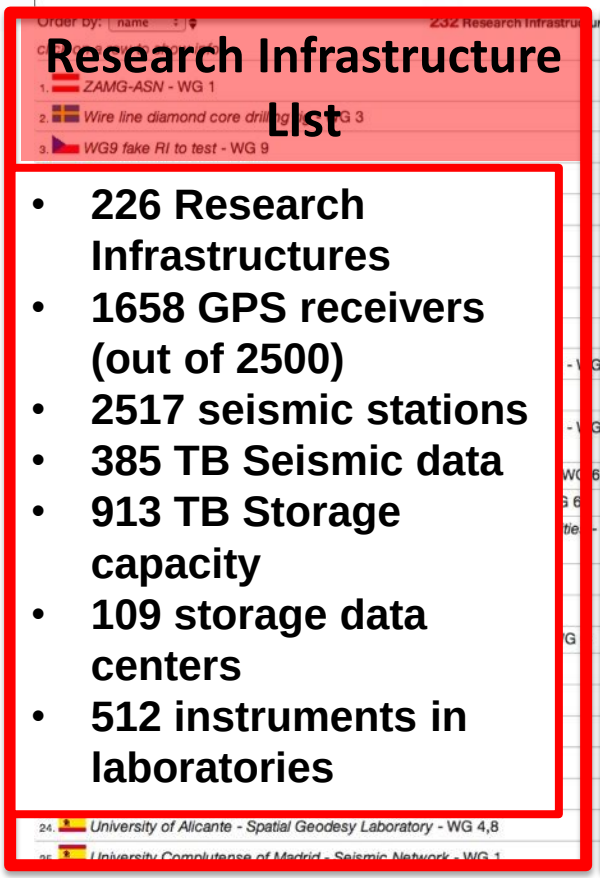
EPOS Framework



EPOS Contents

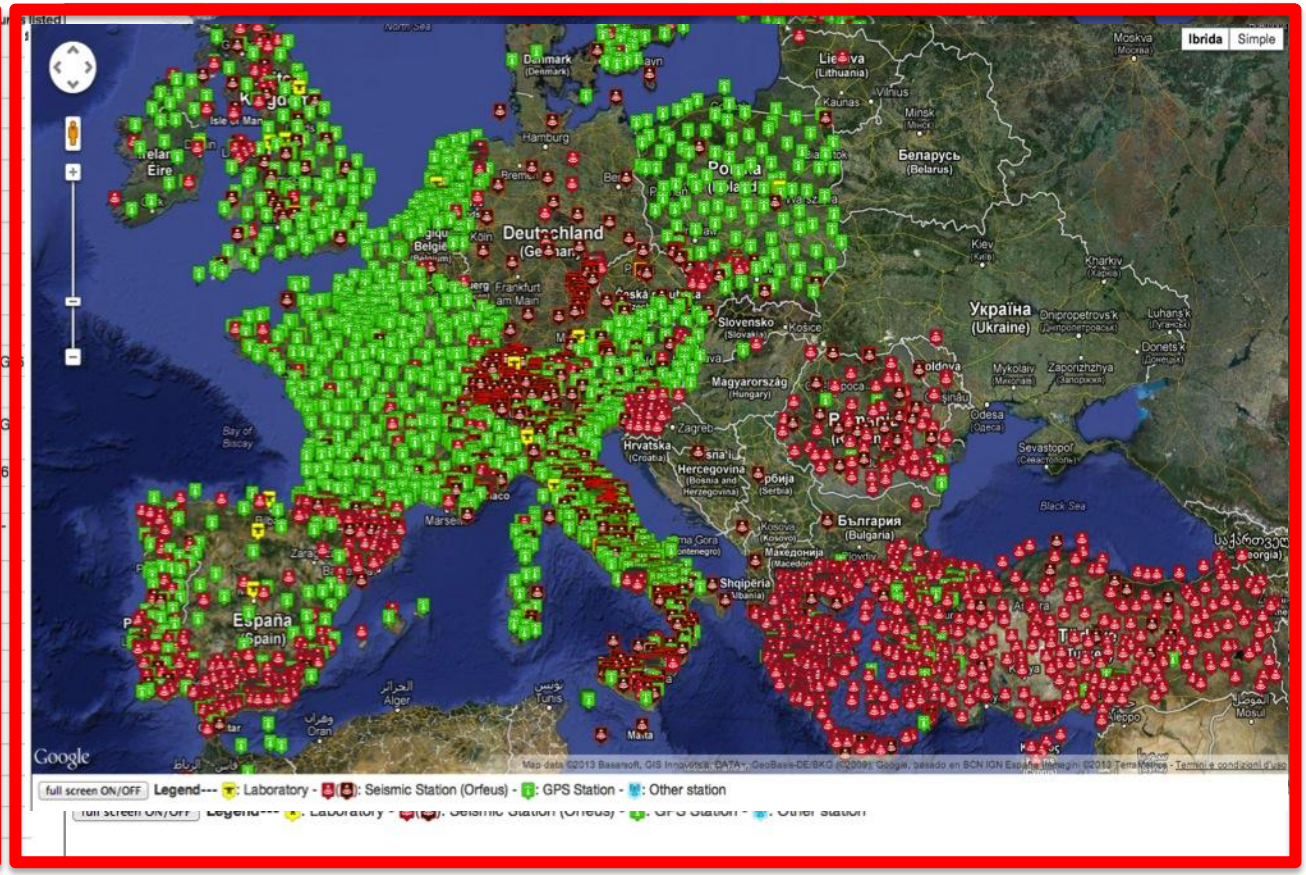
http://www.epos-eu.org/ride/

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



Research Infrastructure List

- 226 Research Infrastructures
- 1658 GPS receivers (out of 2500)
- 2517 seismic stations
- 385 TB Seismic data
- 913 TB Storage capacity
- 109 storage data centers
- 512 instruments in laboratories



Example of EPOS research use case

***Goal*:**

realistic prediction of ground motion in a particular area based on available data and models

Envisaged Steps:

① Discover *largest earthquakes* in the area

From recent and historical catalogues

② Retrieve *moment tensors* (MT) of the earthquakes in 1.

③ Retrieve *finite fault* (if available) or extrapolate the fault finiteness using the available relationships between magnitude, mechanism and fault width, length and slip.

Envisaged Steps (cont'd):

④ Retrieve macro seismic fields for the earthquakes in 1.

⑤ Retrieve *shakemaps* for the earthquakes in 1. for the different PGMs

⑥ Retrieve *velocity structure*

⑦ Retrieve *geologic map* of the target area

⑧ Visualize model+geologic map+hypocenters using interactive 3D graphics

⑨ Plot the *available waveform data*

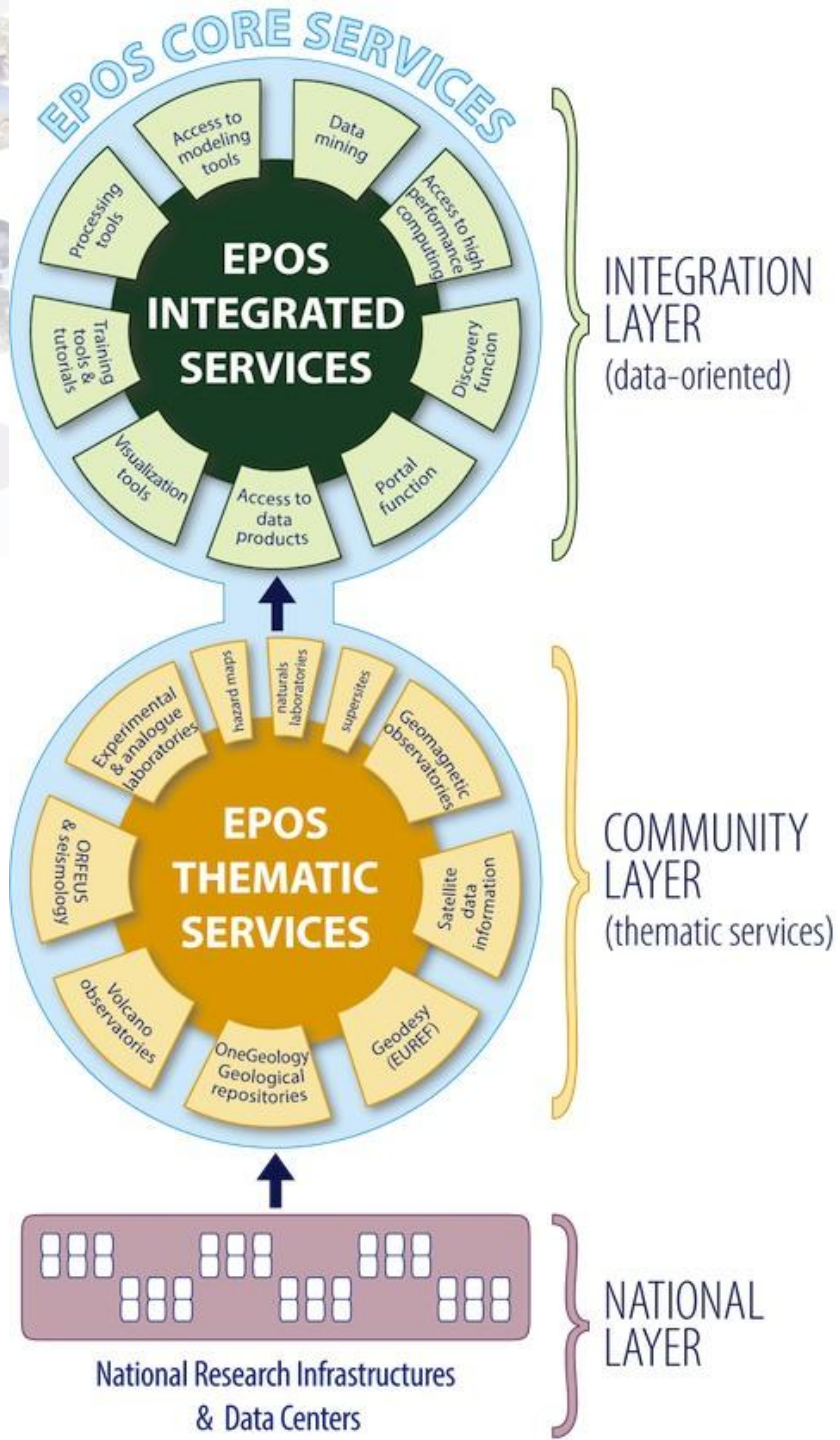
⑩ Simulate waveforms (forward modeling) for the earthquakes in 1. and MTs in 2.

11 Calculate misfit between *observed* and *calculated* waveforms

12 Modify *velocity model* and redo steps 9. and 10.

13 If OK match between observed and synthetics, plot the PGMs on a map of the area

14 Compare *calculated* and *observed ground motion*.



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

EPOS Board of Service Providers

Thematic Services: an example from seismology

EPOS Seismology Products and Services (ESPS)

Governance and coordination by Board of Service representatives, 4-6 members

WAVEFORM DATA

Ground motion recordings from seismic sensors *(possible extension to infrasound)*

Structure: Distributed (ORFEUS umbrella)
~8 nodes, including ORFEUS & EIDA nodes, SISMOS, SMdB

Products (indicative list)
Continuous and event waveforms from permanent and temporary stations (broadband, short period, strong motion); historical waveform archive; synthetic waveform data; strong motion data (products)

Services (…)
Station information (metadata, site characterization…); data quality (control) information

European Infrastructures
Mobile pools, OBS pools…

EARTHQUAKE PRODUCTS

Parametric earthquake information and event-related additional information

Structure: Distributed
~ 5 nodes, including EMSC & its key nodes, AHEAD

Products (indicative list)
Earthquake parameters & bulletins; earthquake catalogues (instrumental, macroseismic, historic, synthetic); moment tensors; source models

Services (…)
Rapid earthquake information dissemination (felt maps, ShakeMaps)

HAZARD AND RISK

Seismic hazard & risk products and services

Structure: Distributed
~3 nodes, including EFEHR (EUCENTER & ETH nodes)

Products (indicative list)
Hazard: Fault maps & models; source zones; hazard maps & curves & disaggregation; GMPEs
Risk: Inventories & inventory models; vulnerability functions; risk maps & scenarios

Services (…)
Tools for model building and visualization; product viewer; hazard & risk calculation software & infrastructure

COMPUTATIONAL SEISMOLOGY

High performance and high end computing, data intensive computing

Structure: Distributed
~3 nodes (build upon VERCE)

Products (indicative list)
Tools for massive scale data applications (processing, mining, visualization,…)

Services (…)
Access to HPC resources; data staging; data massive applications; data simulation; model repository and model handling tools (large 3D velocity models, rupture models,…)

EPOS Volcanology

EPOS Geology

… Other EPOS Communities

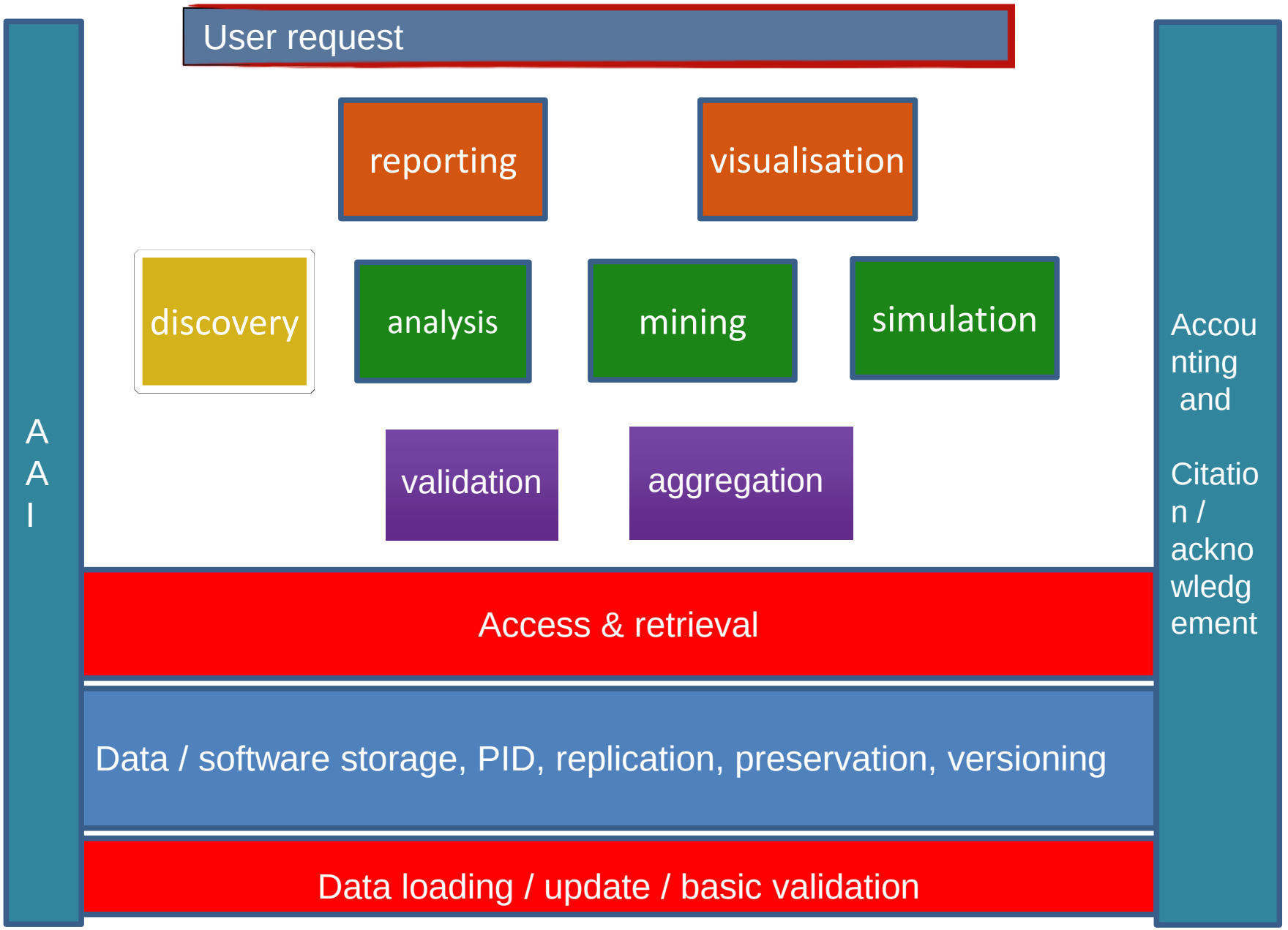
e-Seismology & common services

Seismological services for visualisation, discovery and access to portal (based on seismicportal.eu)
expert groups, standards

EPOS Integrated Services

Visualisation tool / discovery & access portal
high performance and high end computing (may absorb E-Seismology)
expert groups, standards

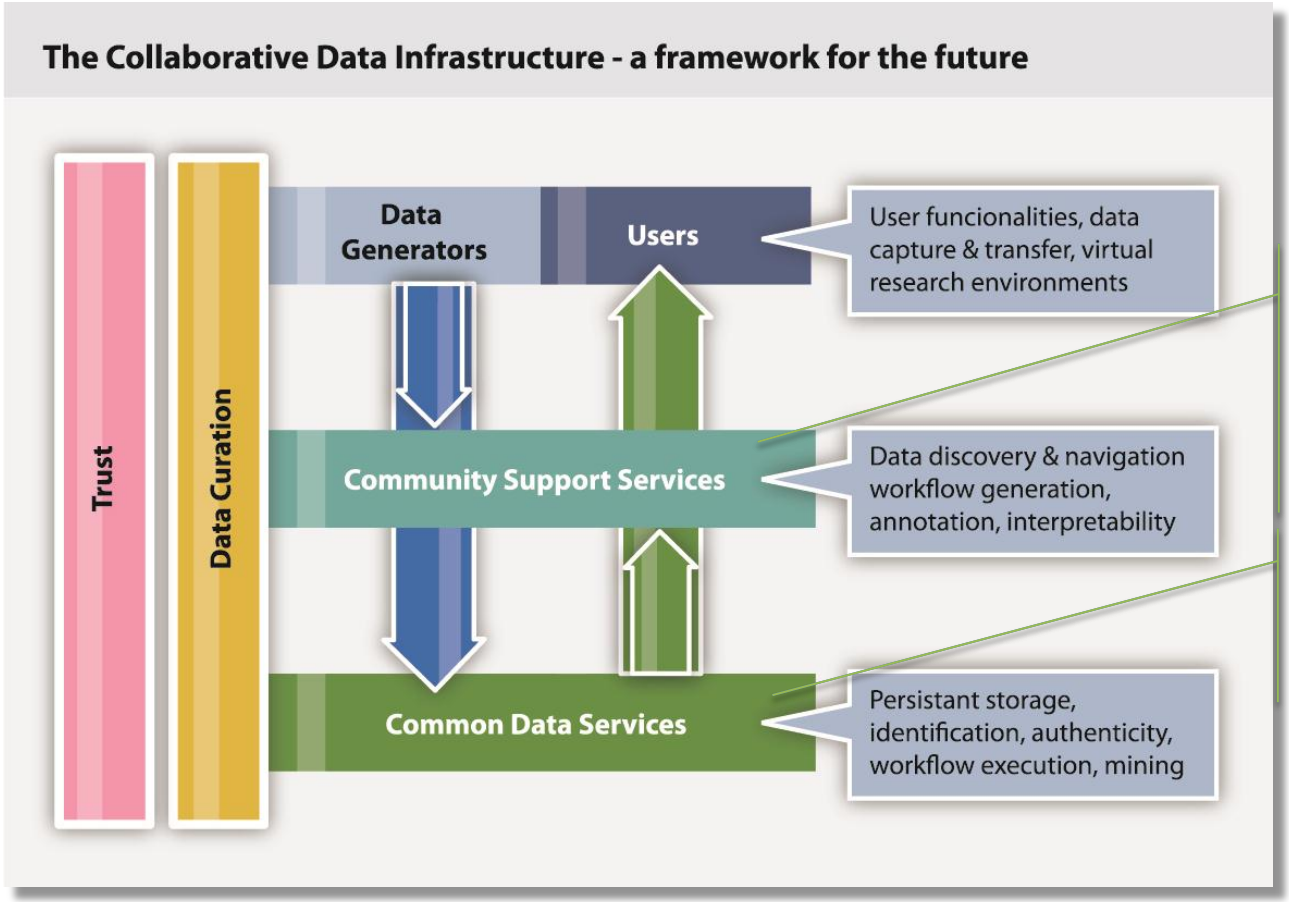
EPOS e-infrastructure model



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**

EUDAT's mission: common services in CDI



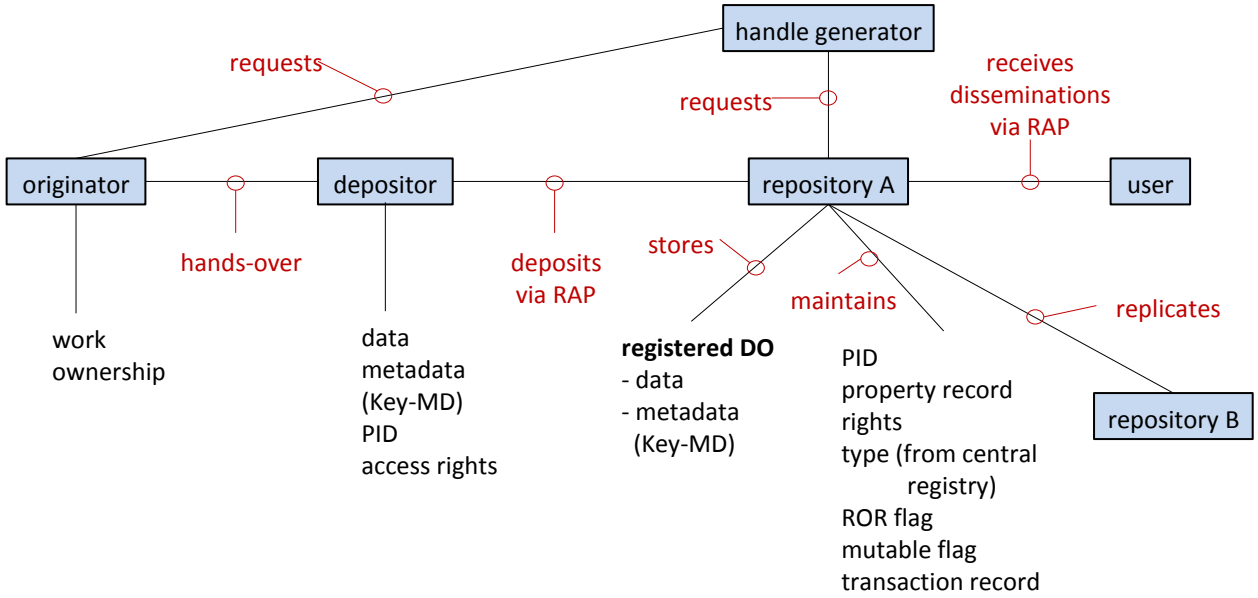
CLARIN, LifeWatch, ENES, EPOS, VPH, etc.
5 Core Infrastructures more second round infrastructures

=> 12 EUDAT data centers

EUDAT

Data Organization and Terminology

- community interactions based on abstract model (Kahn & Wilensky, 2006)
  - Data + Metadata + Handle (PID)
- used in many meetings and interactions - accepted quickly as reference model
- helped even in improving community organization plans

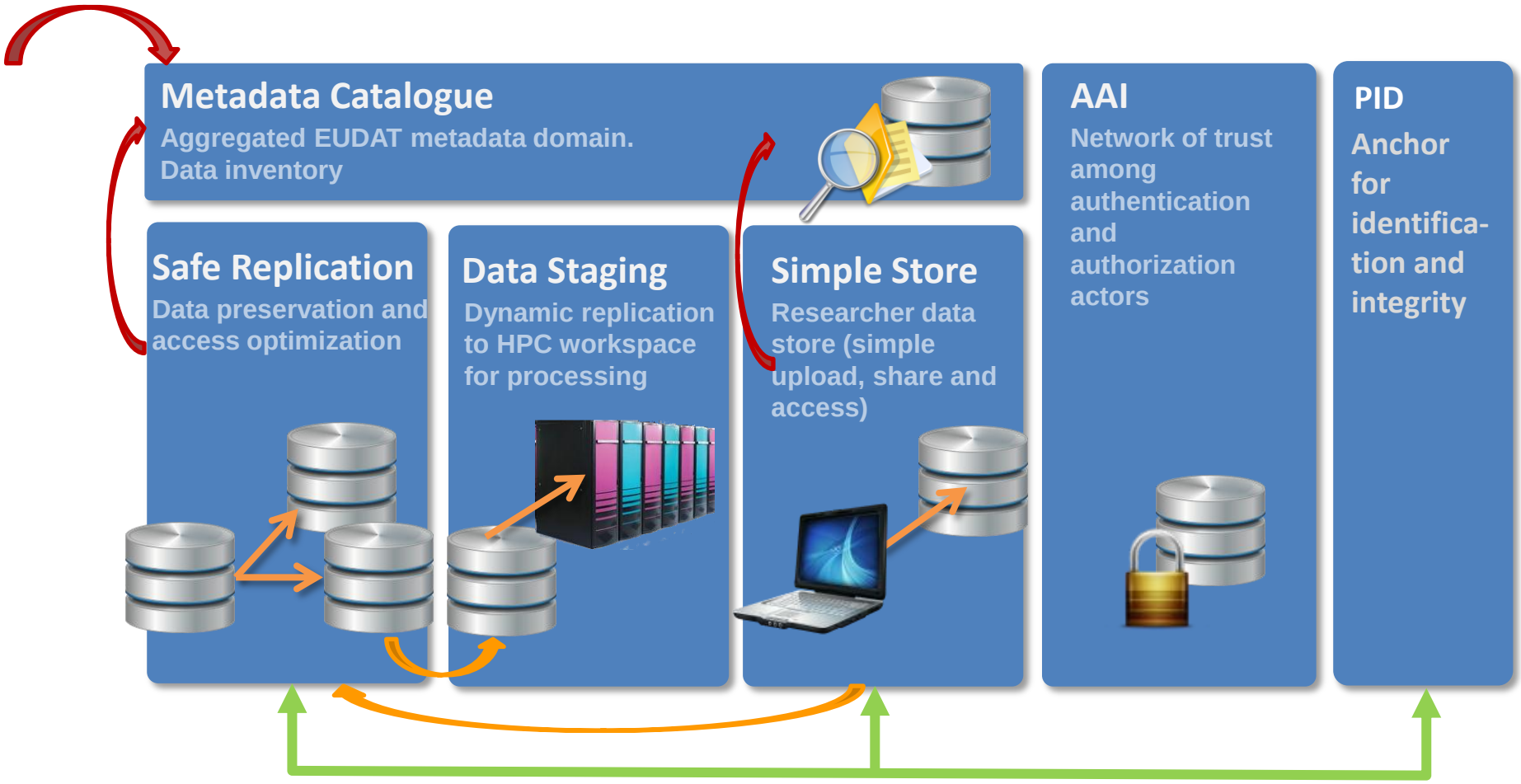


Definitions/Entities

originator = creates digital works and is owner;
depositor = forms work into DO (incl. metadata),
digital object (DO) = instance of an abstract data type;
registered DOs are such DOs with a Handle;
repository (Rep) = network accessible storage to store DOs;
RAP (Rep access protocol) = simple access protocol
Dissemination = is the data stream a user receives
ROR (repository of record) = the repository where data was stored first;
Meta-Objects (MO) = are objects with properties
mutable DOs = some DOs can be modified
property record = contains various info about DO
type = data of DOs have a type
transaction record = all disseminations of a DO

EUDAT

First EUDAT Services



EUDAT

Data PID Metadata

Summary

- Individual **communities have their own thematic services** developed throughout many years and, in general, they are happy with them (!) ➜ ***ad hoc* solutions**
- EUDAT is proposing a **data organization model** which can be instrumental toward e-infrastructures
- EUDAT is developing primarily **core services** common to all the e-infrastructures
- EUDAT and VERCE are posing particular attention to **large-to-huge data volumes analysis**
- In EPOS (solid Earth sciences), **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
  (University curricula should make attention to this)

Summary (cont'd)

- 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 ➜ **evidence improvements**.
- To achieve the best results, it needed **continuous orchestration** between scientific communities and ITs (e.g., scalability, AAI)
- The EUDAT participating partners are effectively "**ambassadors**" of their own community and work is done to **disseminate** the project developments **within the communities** of belonging
- 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**

Thank you

Acknowledgments

Massimo Cocco and the EPOS Team.

The EUDAT and VERCE projects.

The INGV data archive and real-time analysis team.

CINECA for helping us to move the first steps in the world of modern data organization and its services from the IT perspective.

Summary

- Individual communities have their own specific services and, in general, they are happy with them
- Data sharing has enormous potential but the feeling is that there is not yet enough expertise on the resulting advantages, e.g., on the science that can be done by “mixing”/correlating different information (Univerisity curricula should point attention to this)

- EUDAT is developing primarily „core services“common to all the e-infrastructures
- Building an e-infrastructure is very demanding given the diversification of the communities in terms of **different levels** of **data organization development/maturity**
- Must capitalize on the existing developemnts in order to avoid to loose pieces (communities) along the way. The true actors are the communities
- The communities have **their own running services**
- The EUDAT participating partners are effectively “**ambassadors**” of