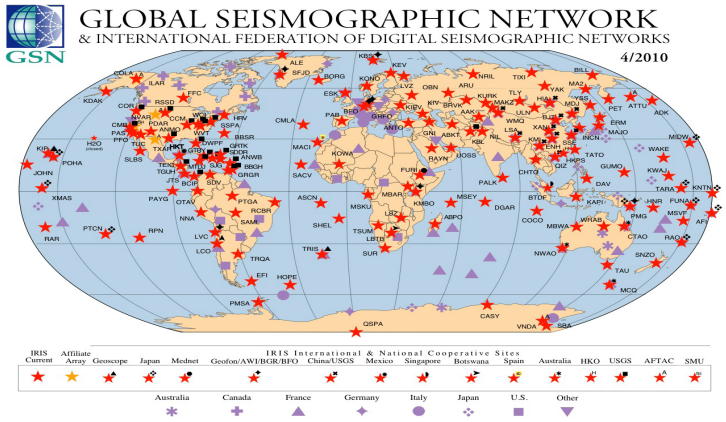
Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Verkeer en Waterstaat

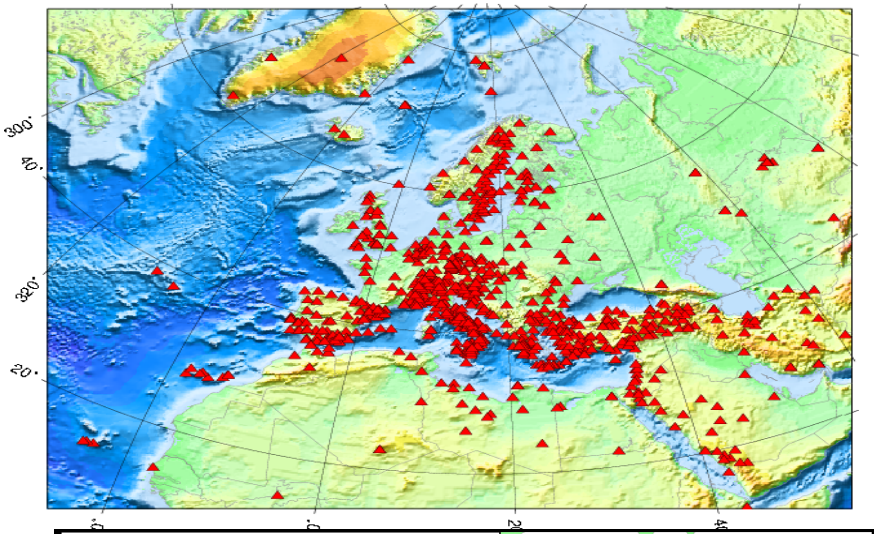
European Infrastructures for Seismology

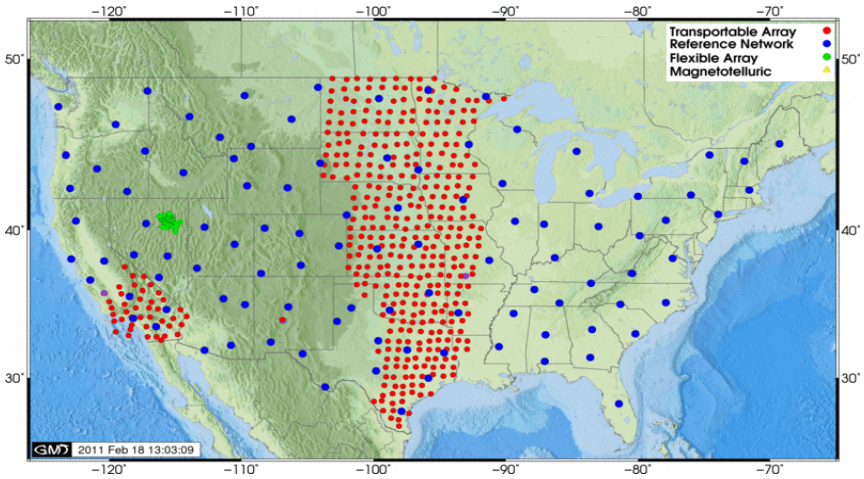
Luca Trani

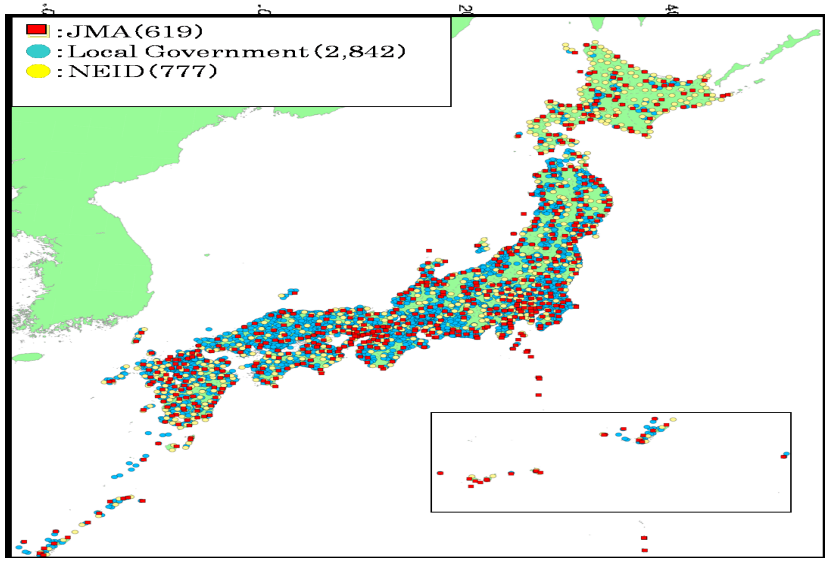
DIR Seminar
19 Oct 2012

A Wealth of Data

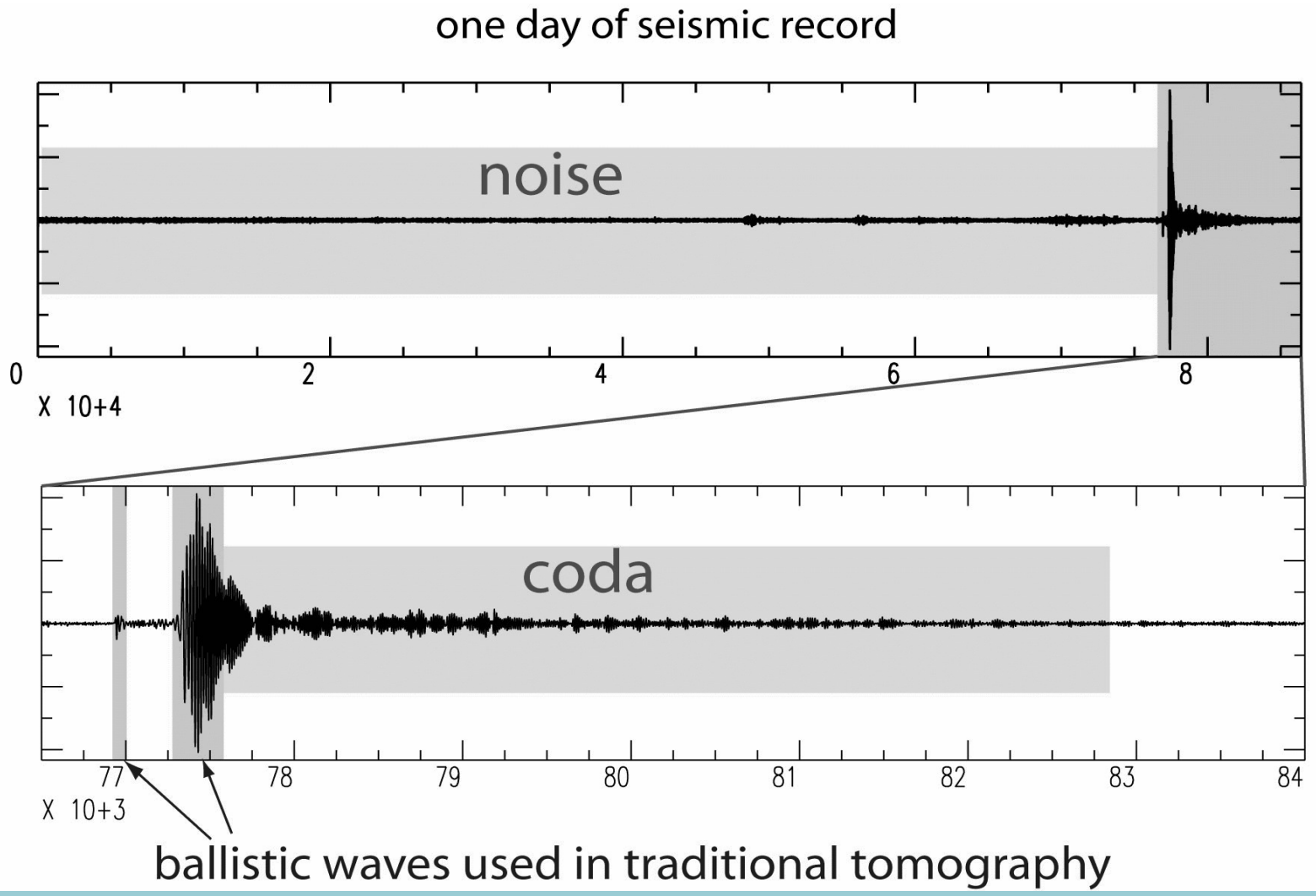




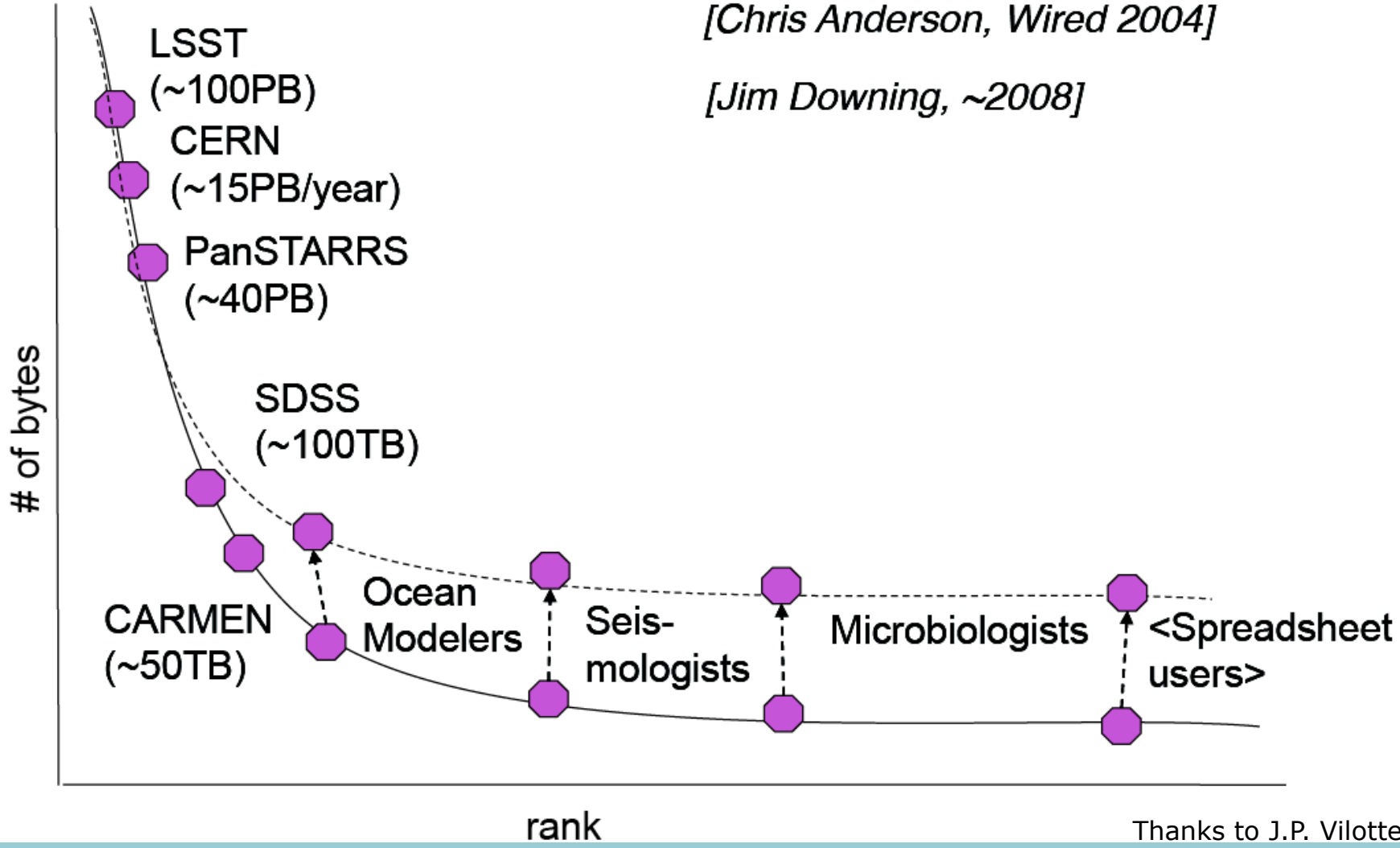




Paradigm shift



The long "Heavy" tail



Thanks to J.P. Vilotte

The need for e-infrastructures

Gigantic Earth Science Data Volumes require the development of new approaches to web-based data and model exchange, data mining and visualization (500 seismometers yield ≈17 GB/day and 6.2 TB/year)

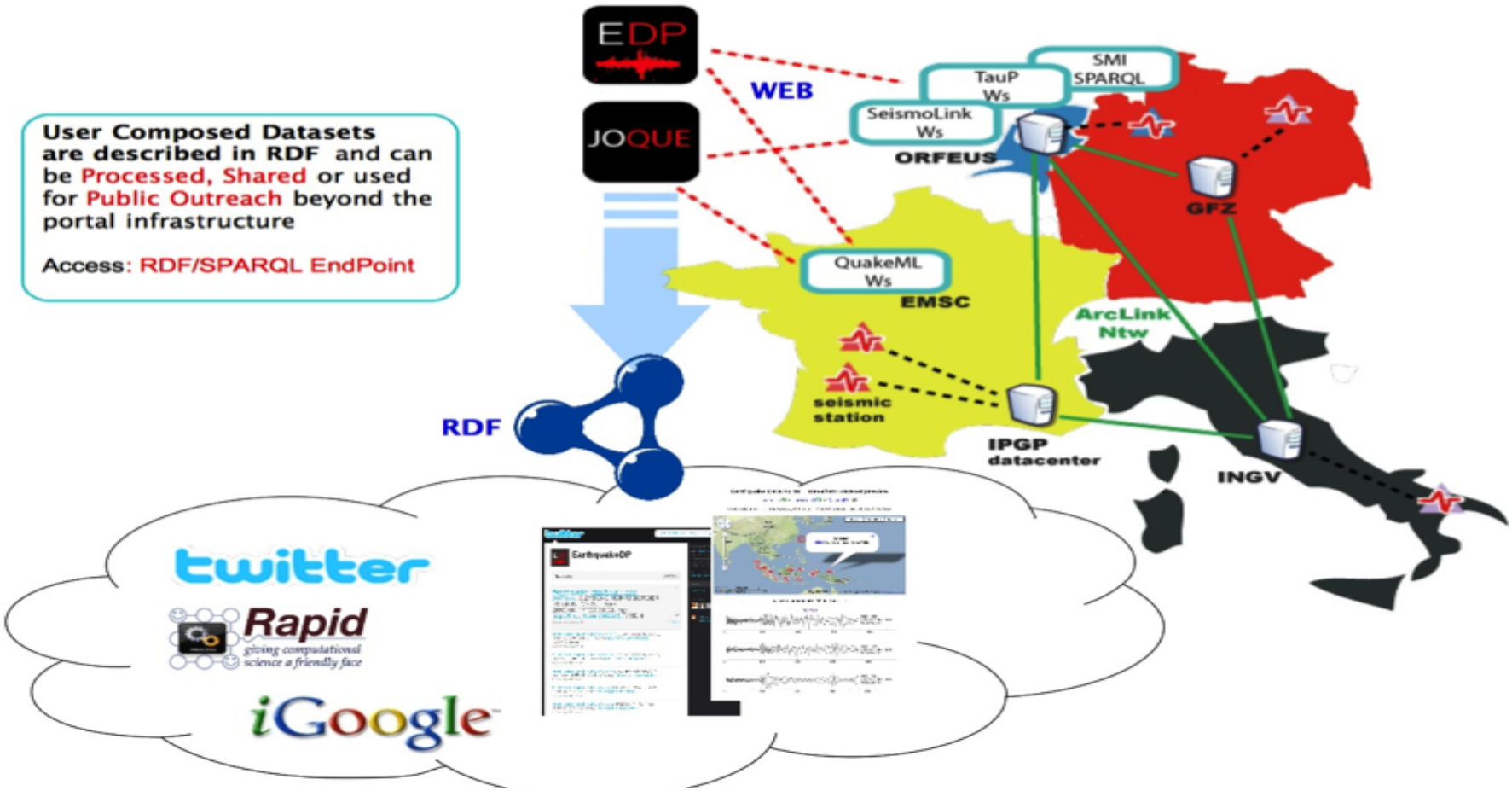
"Virtual Earth Laboratory" - Hypothesis testing will make increasingly use of high-performance simulation technology of Earth's dynamic behaviour

"Software as Infrastructure" – scientific simulation technology needs to be adapted and maintained for wide use by the community

"Data rich" Elements: Web-based superstructure linking Earth Science Data Centres, standardise multi-disciplinary data and model exchange

"Cpu rich" Elements: Simulation and processing technology needs to be professionally engineered, linked to the European High-Performance Computing infrastructure and the scientific data infrastructure

Current infrastructure (thanks to NERIES)



Relevant Projects

NERA
Seismology & Earthq Eng.
ETHZ + ORFEUS/KNMI
(D. Giardini; T. van Eck)

EPOS PP
INGV (Massimo Cocco)

SHARE
Hazard
ETHZ (D. Giardini)

GEM
Hazard

EPOS (ESFRI roadmap)

QUEST (Training network)
Computational Seismology
LMU (H. Igel)

VERCE
IPGP (J-P Vilotte)
ORFEUS/KNMI
EMSC
INGV
LMU
Univ Liverpool
BAW
CINECA
Fraunhofer
UoE (IT)
INFRA-2011-1.2.1

EUDAT
CSC Finland (Kimmo Koski)
EPOS (GFZ, INGV)
LifeWatch
…
CINECA (IT)
UoE (IT)
…
INFRA-2011-1.2.2

ENVRI
LifeWatch (Wouter Los)
EPOS (ORFEUS/KNMI)
LifeWatch
EPOS
EMSO
EISCAT
ICOS
…
STFC (IT)
UoE (IT)
…
INFRA-2011-2.3.3

Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation

Virtual Earthquake and seismology Research Community e-science environment in Europe

Research Infrastructure and E-Science for Data and Observatories on Earthquakes, Volcanoes, Surface Dynamics and Tectonics

European Data Infrastructure

NERA
Real-Time & Rapid Data analysis

VERCE
Platform for Data- and CPU-intensive applications

Seismological

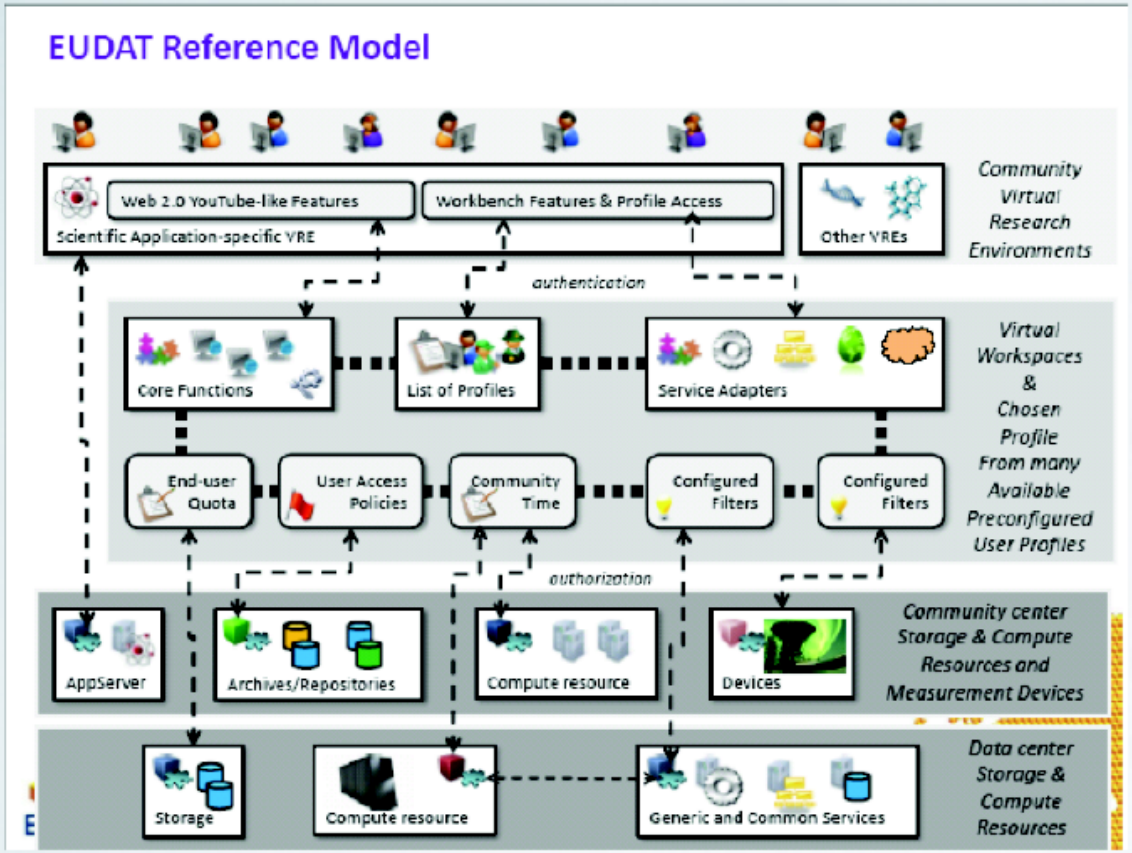
EPOS
EUROPEANPLATEOBSERVINGSYSTEM
Solid Earth e-infrastructure

Solid Earth Sciences

EUDAT
Core Services for e-infrastructures

Data IT

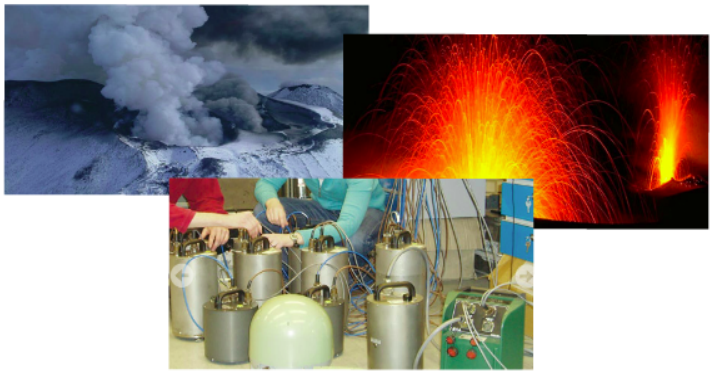
Projects interactions



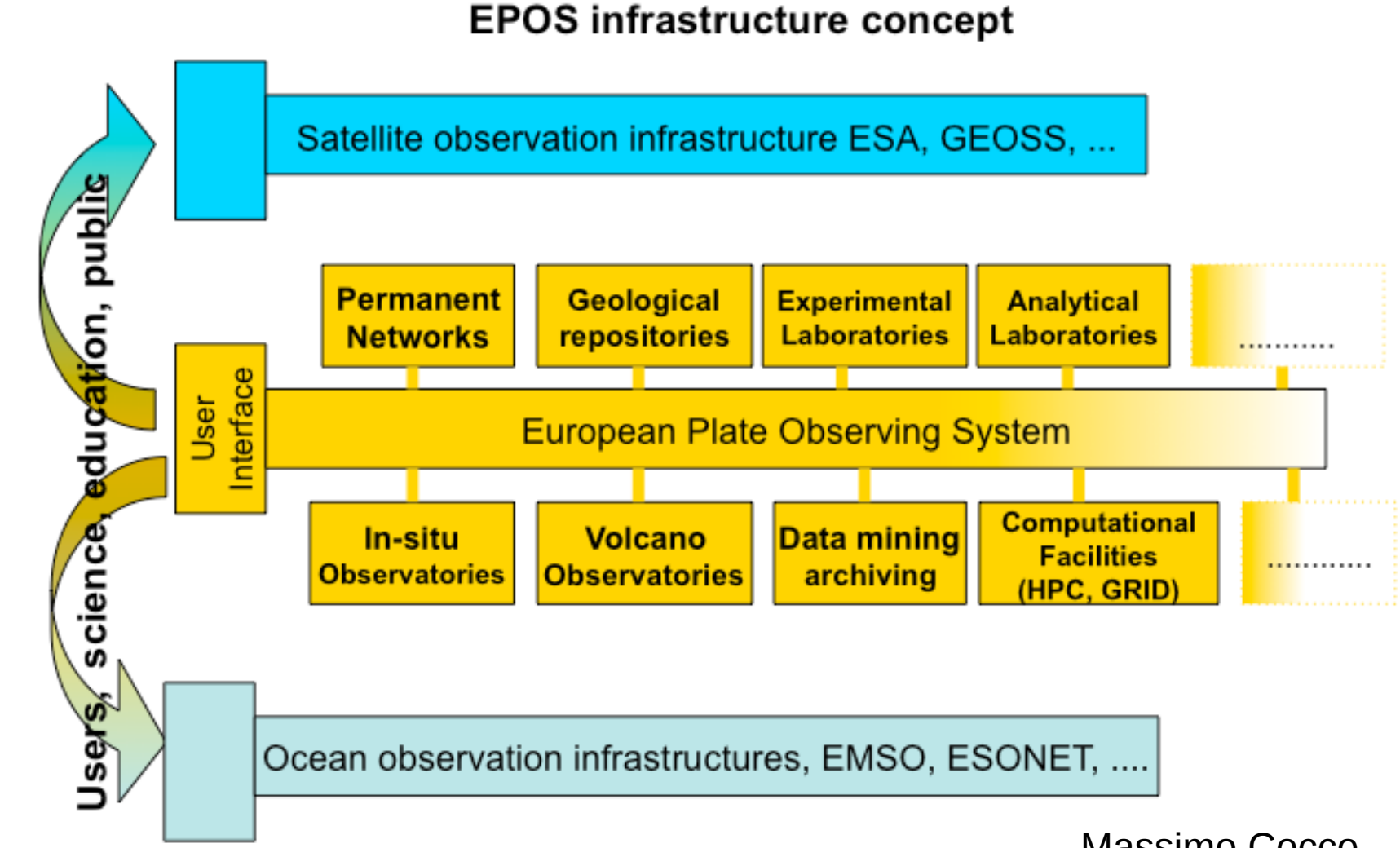
http://www.epos-eu.org/

EPOS - European Plate Observatory System

- Distributed data sensors
- Large scale statistics
- Metadata schema
- Reference architecture



Research Infrastructure and E-Science for Data and Observatories on Earthquakes, Volcanoes, Surface Dynamics and Tectonics



Massimo Cocco

EPOS Services

Data collection, calibration, validation

Data cataloguing and indexing

Data preservation and curation

Information processing – retrieval, analysis, visualisation

Hypothesis processing – simulation, modelling, analysis, visualisation

Hypothesis generation – data mining

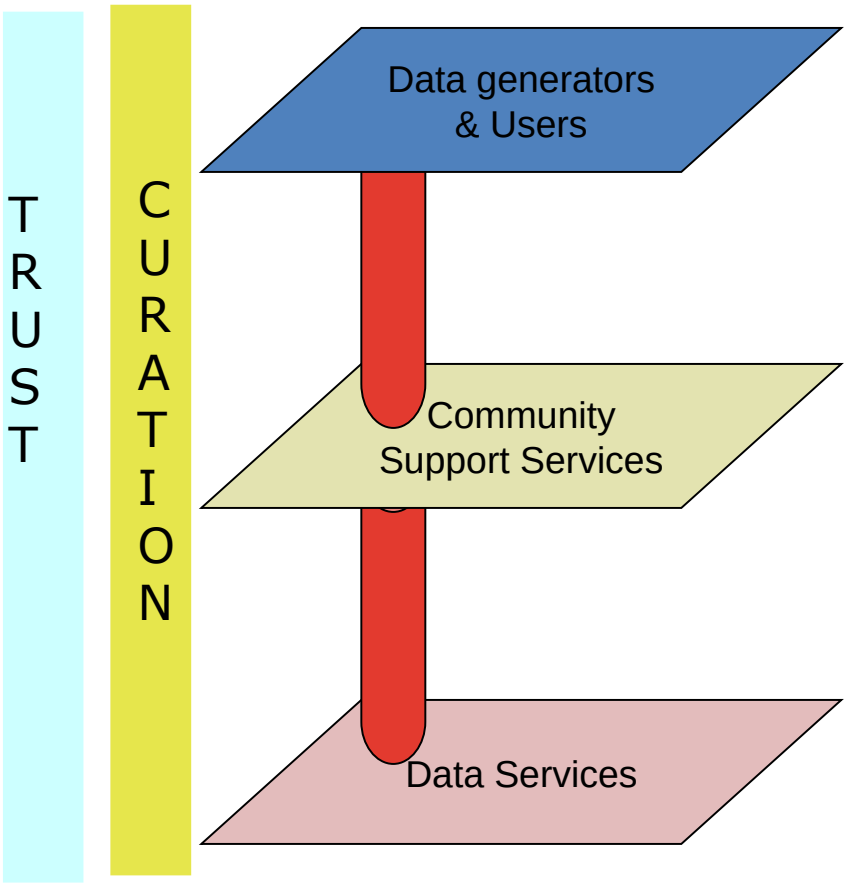
Knowledge processing – integration of ICT with human processing – theory processing, user interface, scholarly communication

- External interoperation – physical and medical sciences, economic and social sciences, arts and humanitiesDissemination – outreach

Education and training

Management and Coordination

The Eudat CDI concept



Data providers & Users Humans & Instruments

Roles

Sensors Curators Researchers

Observers Aggregators Public

Functionalities

Virtual Environments & Collaborative organisations

Security & Protection

Data discovery & Navigation

(meta) data tagging tools

Data submission tools

Operational Semantic Interoperability

Workflow Generator

Data correlation

Knowledge management

Virtualisation

Persistent storage capacity

24/7 operation

Preservation & Sustainability

Authenticity

Certification & Integrity

GUIDs

Generic interoperability

Technical

Legal

Semantic

EUDAT Core Services

Community-oriented

Enabling Services

EUDAT Portal

Integrated APIs and harmonized access to EUDAT facilities

Metadata Catalog

Aggregated EUDAT metadata domain.
Data inventory

Requirement
Provide an inventory of metadata across disciplines

Function
- Joint metadata domain
- Catalog indexing stored data

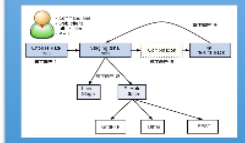
eudat-metadata@postit.csc.fi

PID-based index

Data Staging

Dynamic replication to HPC worspace for processing

Requirement
Provide a service to stage data between EUDAT infrastructure and HPC/HTC resources



Function
Dynamically replicate subset of data stored in EUDAT to HPC workspace

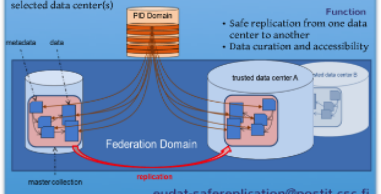
eudat-datastaging@postit.csc.fi

Safe Replication

Data curation and access optimization

Requirement
Provide a service to replicate and curate data to selected data center(s)

Function
- Safe replication from one data center to another
- Data curation and accessibility

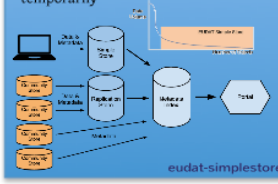


eudat-safereplication@postit.csc.fi

Simple Store

Researcher data store (simple upload, share and access)

Requirement
Provide a simple service to store user data temporarily



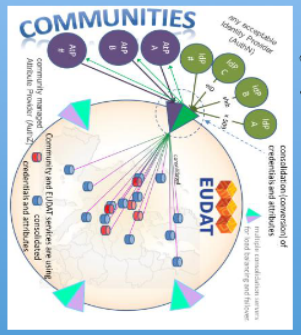
Function
Simple upload
Store data

eudat-simplestore@postit.csc.fi

AAI

Network of trust among authentication and authorization actors

Requirement
Provide a working AAI system in a federated scenario



Function
- Integrate existing identification systems
- Establish a network of trust among AA, IdP and SP providers, attribute authorities and federations attribute harmonization

eudat-AAI@postit.csc.fi

Building Blocks of the Collaborative Data Infrastructure

http://www.eudat.eu/services-and-technologies

Safe Replication @ EUDAT

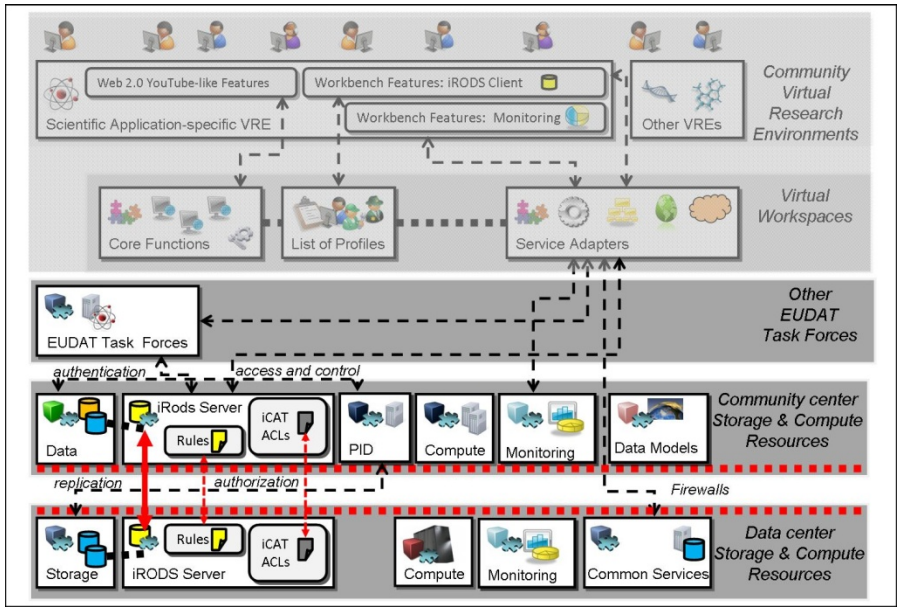
Objective: Enable communities to easily replicate data to selected data centres for storage in a robust and reliable manner.

Key benefits: data bit stream preservation, more optimal data curation, better accessibility

Description: Data replication management based on users' requirements and constraints; data replication solutions and services embedded into critical security policies, including firewall setups and user accounting procedures.

Technology: iRODS to be used as an initial replication middleware, implemented across the community centres and data centres; as more user communities join the task force, other storage technologies may be added, depending on user needs.

- Production setup expected by 2013, such that users will be able to safely replicate data across different user community centres and data centres.



i·R·O·D·S
Integrated Rule-Oriented Data System

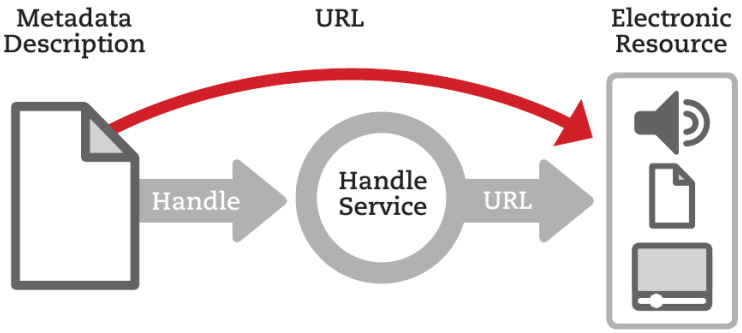
More info: eudat-safereplication@postit.csc.fi

PIDs@EUDAT

Objective: Deploy a robust, highly available and effective PID service that can be used within the communities and by EUDAT.

Description: Keeping track of the "names" of data sets or other digital artefacts deposited with the CDI requires more robust mechanisms than "noting down the filename". The PID service will be required by many other CDI services, from Data Movement to Search and Query.

Technologies: Currently considering use of both EPIC for data objects, and DataCite to register DOIs (Digital Object Identifiers) for published collections.



More info: eudat-persistentidentifiers@postit.csc.fi

Data Staging @ EUDAT

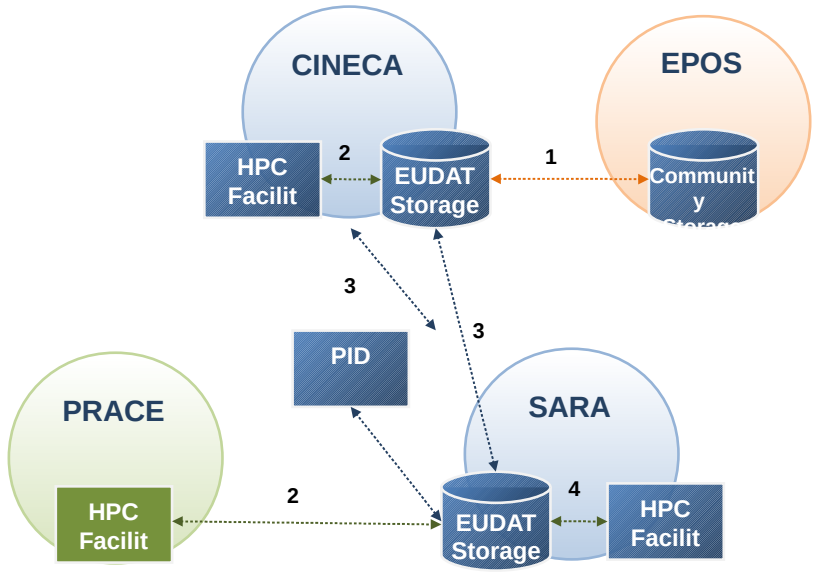
Objective: Enable communities to perform (HPC) computations on the replicated data

Key benefits: Access to large computing facilities

Description*:* This service will allow the EUDAT communities to dynamically replicate subsets of their data stored in EUDAT to HPC machine workspaces for processing.
Differences with the safe replication scenario:

- replicated data are discarded when the analysis application ends;
- Persistent Identifier (PID) references are not applied to replicated data into HPC workspaces;
- Users initiate the process of replicating data while in the safe replication scenario data are replicated automatically on a policy basis.

Technologies: GridFTP, Griffin, gTransfer, FTS (under appraisal)



More info: eudat-datastaging@postit.csc.fi

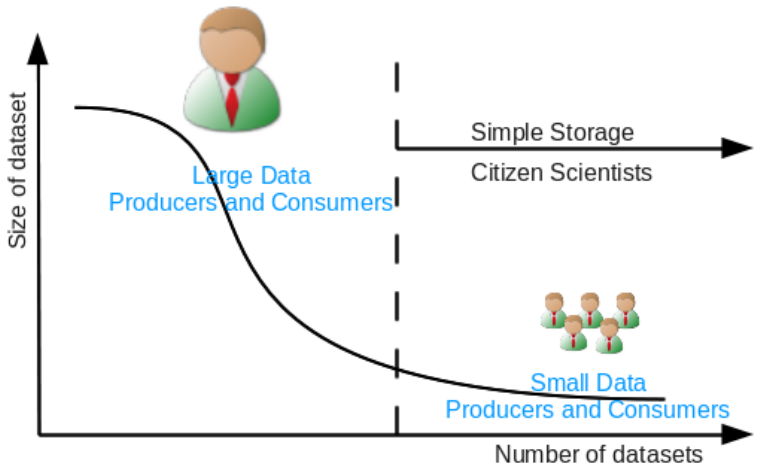
Simple Store @EUDAT

Objective: create an easy-to-use service that will enable researchers and scientists to upload, store and share data that are not part of the officially-managed data sets of the research communities.

Key benefits: Store, share, and retrieve smaller sets of data not officially handled.

Description: This service will address the long tail of "small" data, and the researchers/citizen scientists creating and manipulating it. Typically this type of data comes in a wide range of formats including text, spreadsheets, number series, audio and video files, photographs and other images. The Research Data Store is complementary to the other EUDAT services that manage the large volumes of official community data.

Technologies: Invenio, figshare, beehub and MyExperiment.



More info: eudat-simplestore@postit.csc.fi

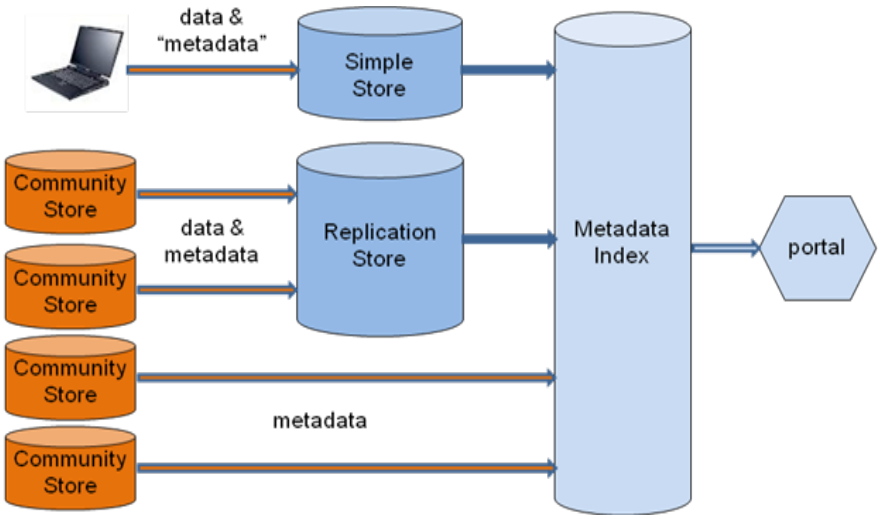
Metadata@EUDAT

Objective: Create a joint metadata domain for all data stored by EUDAT data centres and a catalogue which exposes the data stored within EUDAT, allowing data searches.

Key benefits: Advertising platform for data sets, metadata service for less mature communities

Description: EUDAT will handle metadata for more resources than just those deposited within the EUDAT CDI. In the initial phase we will target mainly resources contributed by the participating communities augmented with those of interested well-organized communities that are ready to contribute. Then, later, other interested communities can be approached depending on the respective community capabilities.

Technology: OAI-PMH and embeds domain specific metadata, as XML, within the OAI-PMH record



More info: eudat-metadata@postit.csc.fi

AAI@EUDAT

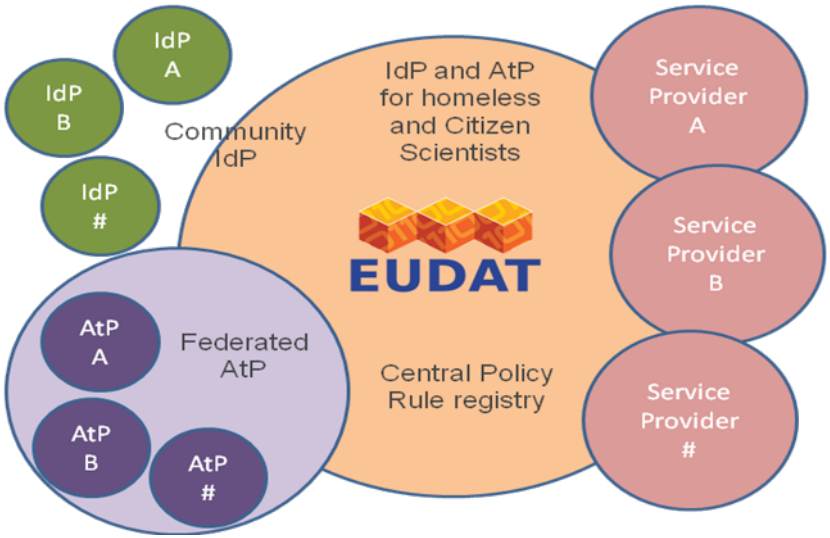
Objective: Provide a solution for a working AAI system in a federated scenario.

Description: Design the AA infrastructure to be used during the EUDAT project and beyond.

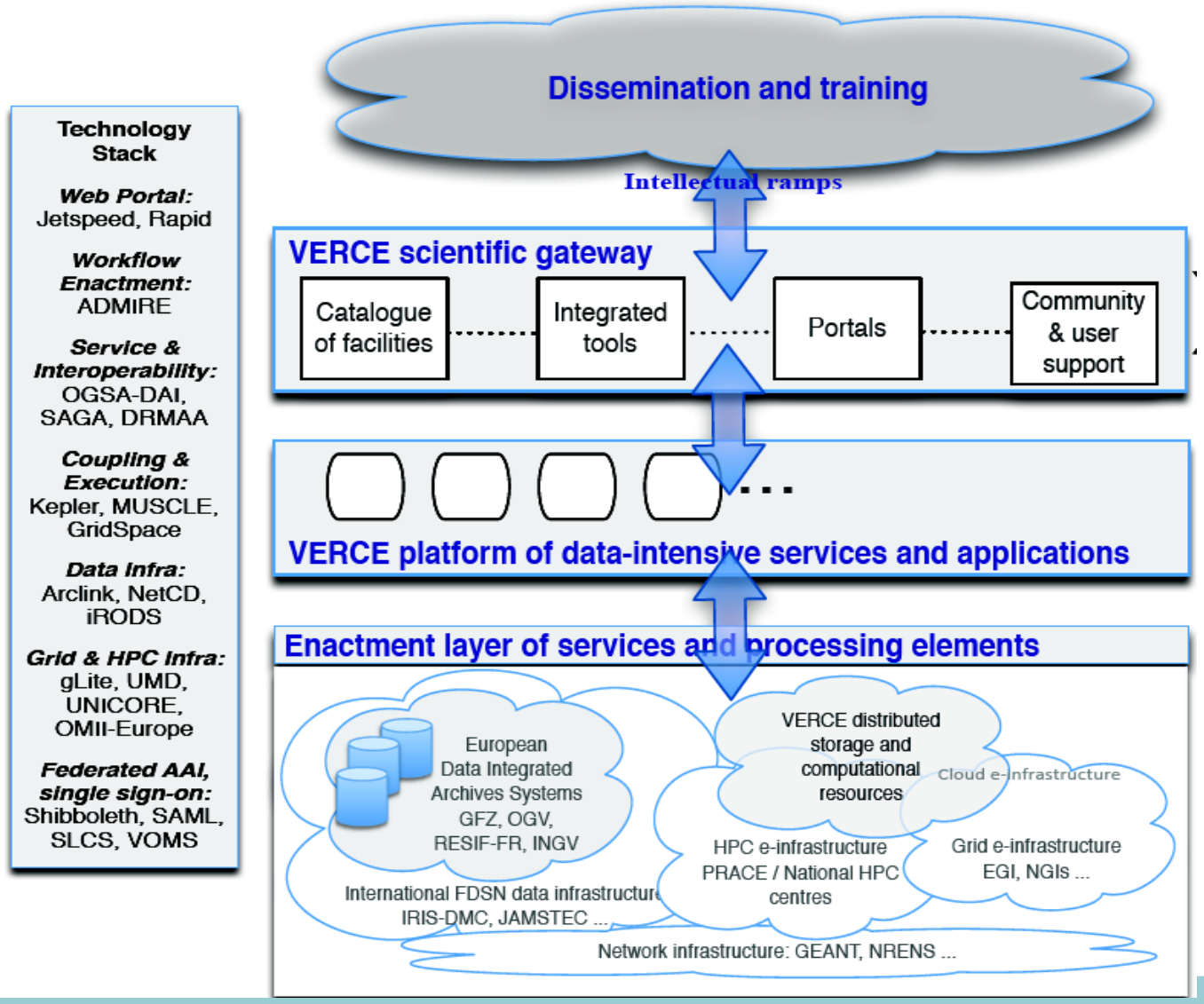
Key tasks:
Leveraging existing identification systems within communities and/or data centres
Establishing a network of trust among the AA actors: Identifty Providers (IdPs), Service Providers (SPs), Attribute Authorities and Federations
Attribute harmonization

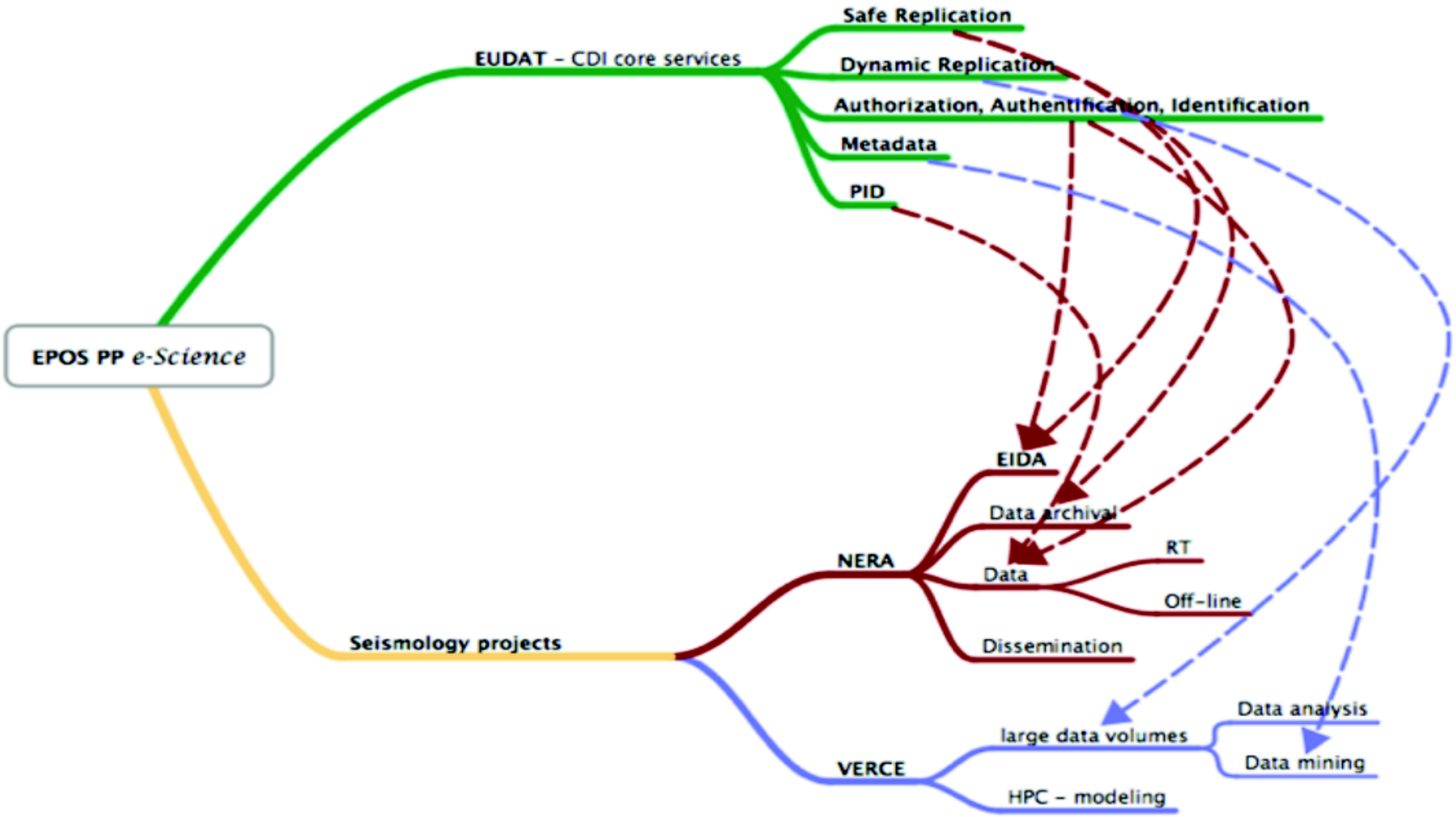
Technologies: Oauth2, OpenID, RADIUS, SAML2, X.509, XACML, etc.

More info: eudat-AAI@postit.csc.fi



The VERCE platform



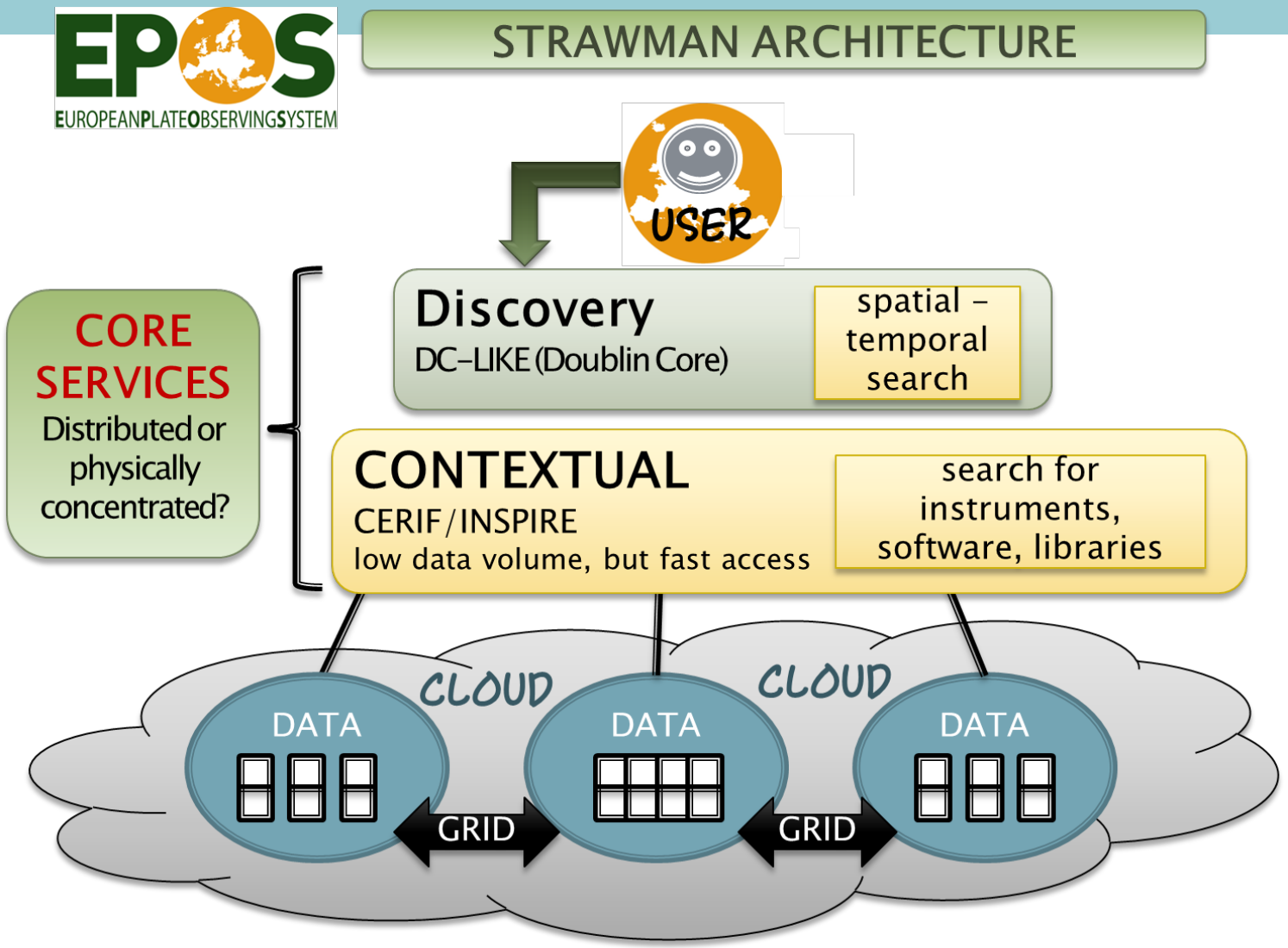

EPOS PP e-Science
EUDAT – CDI core services
Safe Replication
Dynamic Replication
Authorization, Authentification, Identification
Metadata
PID
Seismology projects
NERA
EIDA
Data archival
Data
RT
Off-line
Dissemination
VERCE
large data volumes
Data analysis
Data mining
HPC – modeling

Importance of Metadata

The metadata must be fit for purpose for the following functions related to a digital object:

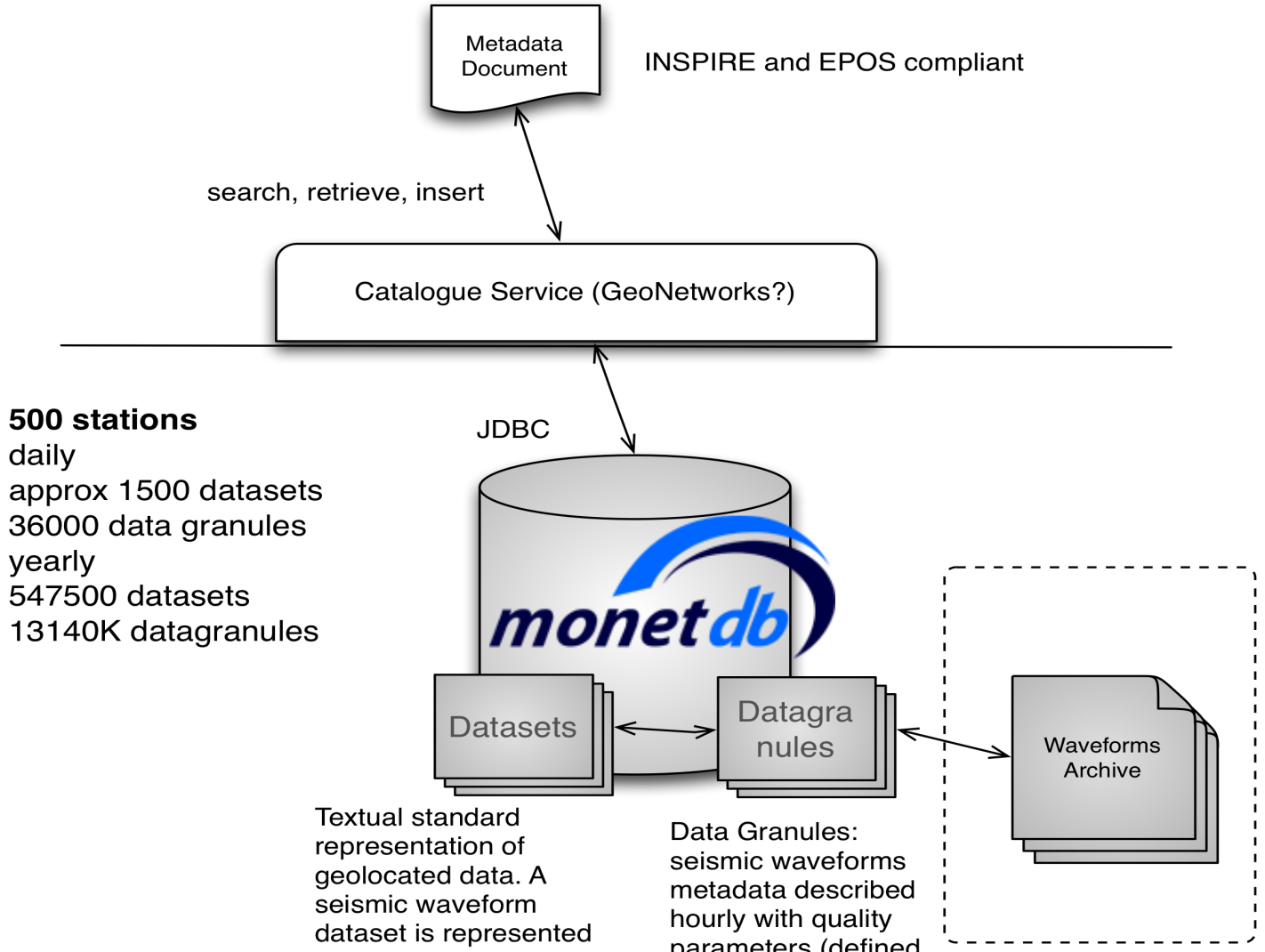
- discovery,
- reading,
- ingesting,
- combining,
- processing,
- outputting as a modified version,
- citing,
- preserving, (all related to organisations, projects, persons, other datasets, publications etc and located in space and time).

Metadata a first design



Data Management @ ORFEUS

NERA



500 stations
daily
approx 1500 datasets
36000 data granules
yearly
547500 datasets
13140K datagranules

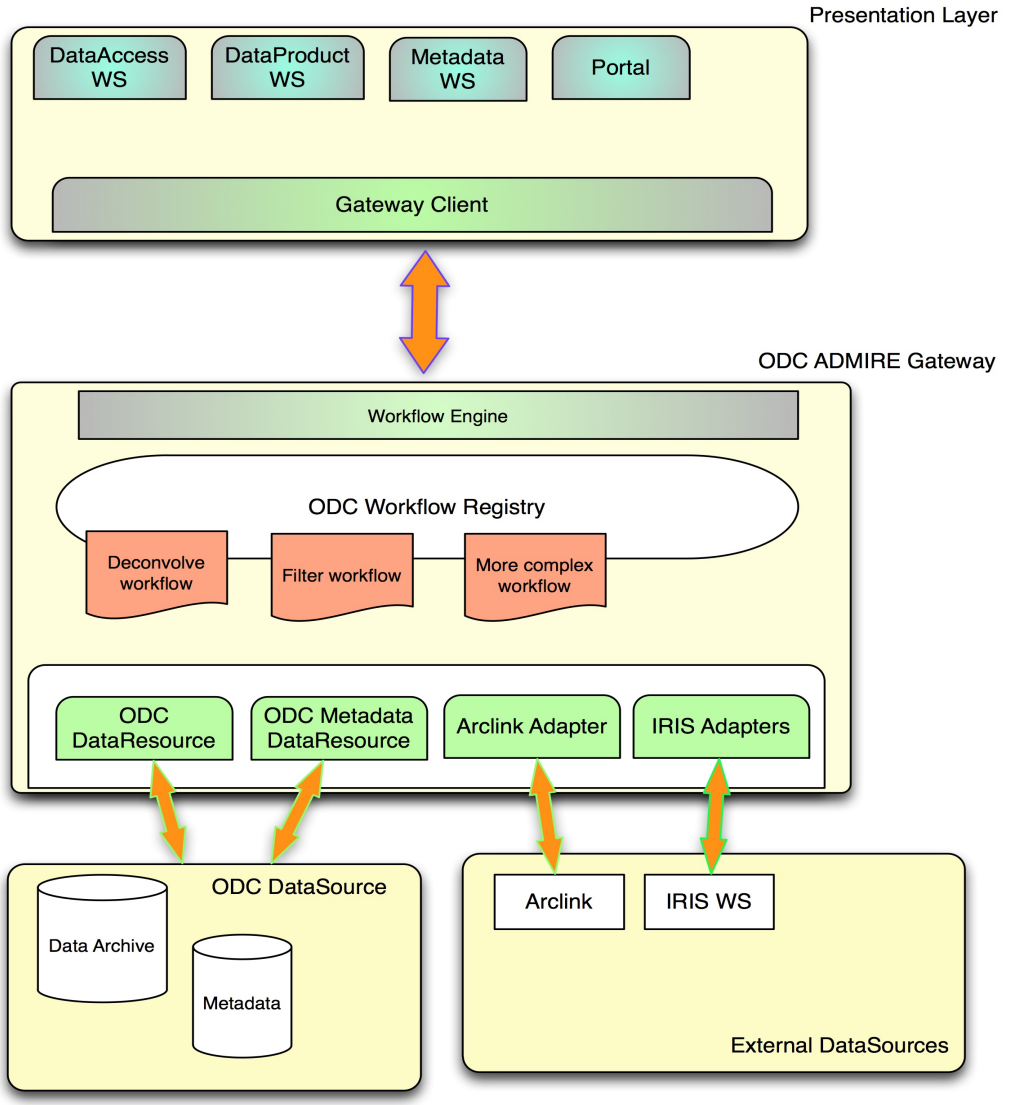
Data Products @ ORFEUS

NERA

Extensible WS Interface for the user services and presentation layer

Data products obtained by workflows execution

Homogeneous extensible collection of adapters(OGSA-DAI activities) for data access



In conclusion

NERA, **VERCE** and **EUDAT** are all EC projects that can provide some major contribution to **EPOS** for the seismological part.

- NERA - should provide some basic services, software and tools eventually integrated in a seismological portal
- VERCE - should provide tools and computational power otherwise not available and, in general, a testbed for implementing and developing HTC and HPC applications.
- EUDAT - should provide expertise and core services for federating and exploiting data located across geographically distributed archives , the availability of user/ groups workspaces and interaction among different users and digital object preservation, curation and access