



# EPOS and the EUDAT CDI

Luca Trani







Data curation and the EUDAT Collaborative Infrastructure Workshop  
Amsterdam 22 feb. 2016



Royal Netherlands  
Meteorological Institute  
*Ministry of Infrastructure and the  
Environment*

[www.eudat.eu](http://www.eudat.eu)

# Outline

-  EPOS Intro
-  Orfeus EIDA
-  EPOS-S Uptake plan in EUDAT2020
-  Working with CDI services
-  Dynamic Data
-  Conclusions

# What is EPOS?

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

EPOS integrates the **existing (and future)** advanced European facilities into **a single, distributed, sustainable infrastructure** taking full advantage of new **e-science opportunities**



Several PetaBytes of solid Earth Science data will be available

Several thousands of users expected to access the infrastructure



# Communities and services

Seismology

Seismic waveforms (ORFEUS)  
Seismological products (EMSC)  
Hazard & risk products (EFEHR)  
Computational seismology

Near fault observatories

NFO multidisciplinary data & products  
Borehole data  
Virtual laboratory & early warning test beds

GNSS data and products

GNSS primary data & derived products  
Processing and visualization tools

Satellite data

SAR interferograms  
Integrated satellite products  
On-line processing tools

Volcano observations

Multidisciplinary volcanic data & products  
Hazard products  
TNA to volcano observatories

Anthropogenic hazards

Data for AH episodes  
Multi-hazard simulator - multi-risk assessment  
AH data visualisation

Geomagnetic observations

Global and regional geomagnetic models  
Magnetotelluric data

Geological information  
and modeling

Geological multi-scale data  
Integrated geological maps  
Borehole visualization

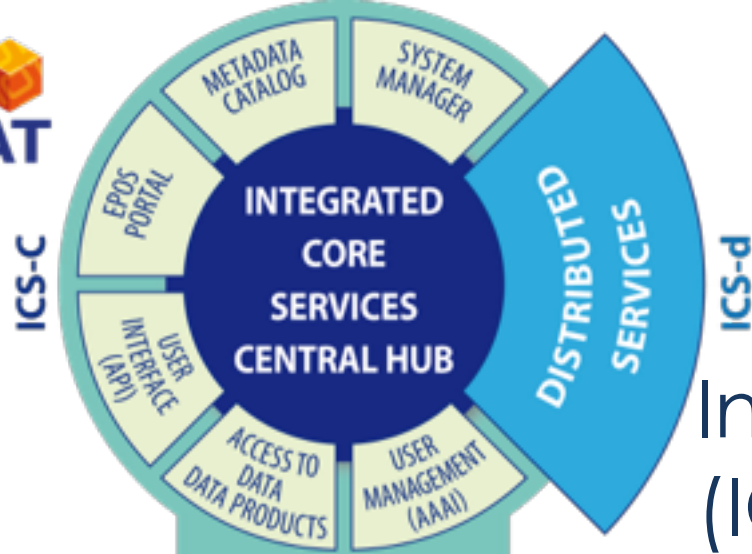
Multi-scale laboratories

Experimental & analogue data  
TNA to experimental & micro-analytical facilities

Geo energy test beds  
for low carbon energy

Geo energy test beds  
Access to in-situ GETB experiments

# Architecture



Integrated Core Services (ICS)



Interoperability layer

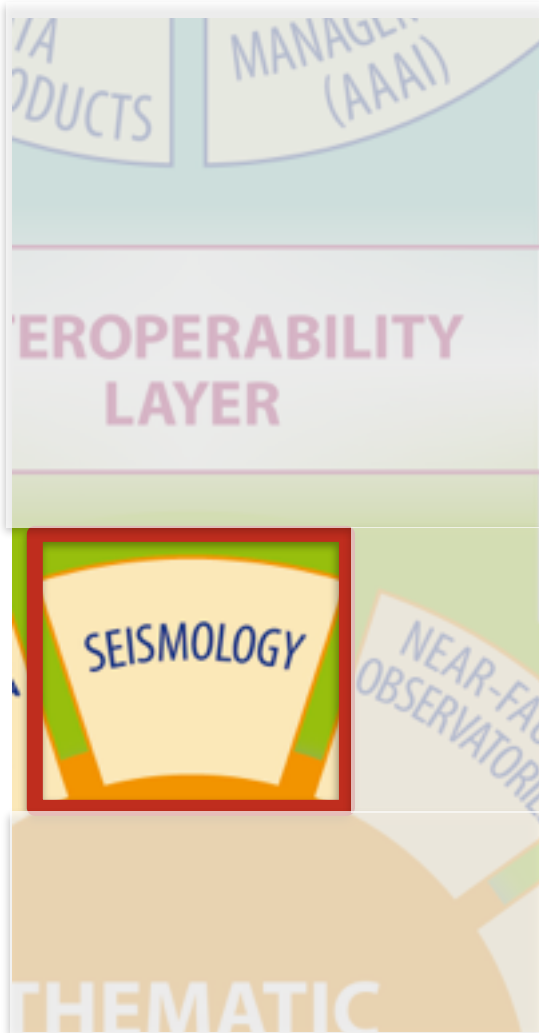


Thematic Core Services (TCS)



NRI

NATIONAL RESEARCH INFRASTRUCTURES & DATA CENTERS



The **E**uropean **I**ntegrated **D**ata **A**rchive is a federated seismological data centre within **Orfeus**.

**Orfeus** coordinates and promotes digital broadband seismology in the European Mediterranean area

[www.orfeus-eu.org/eida/eida.html](http://www.orfeus-eu.org/eida/eida.html)

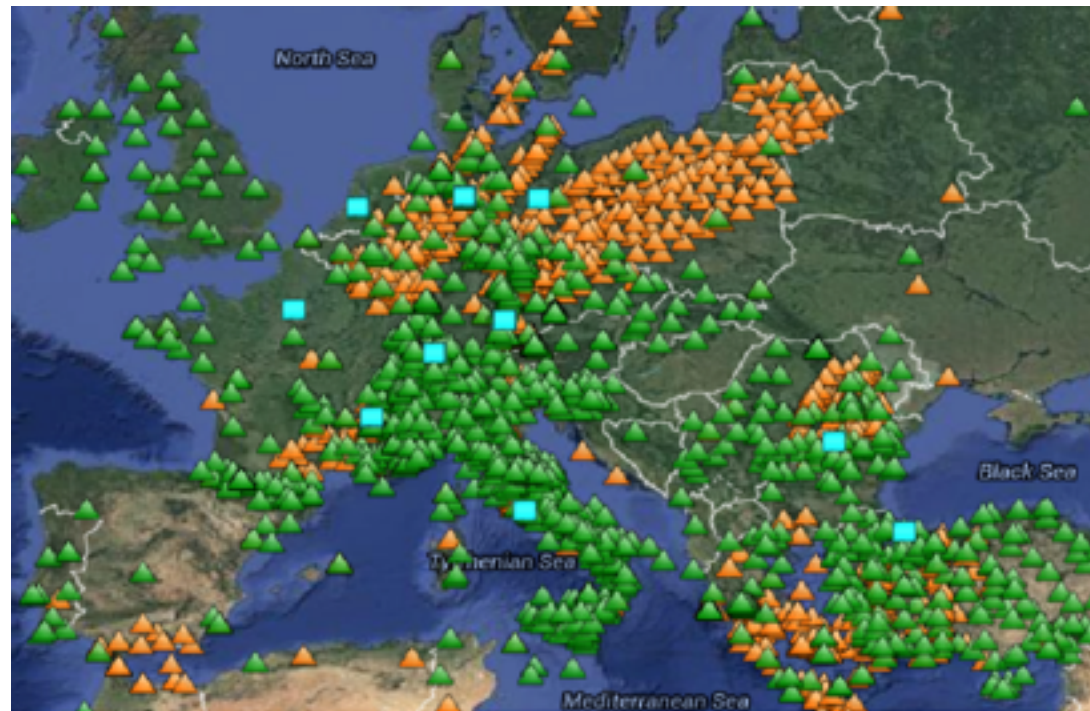


**Continuous** seismic **streams** are collected in real time from contributing networks.

Data are archived in **federated**, geographically distributed data centers which implement data management policies to guarantee:

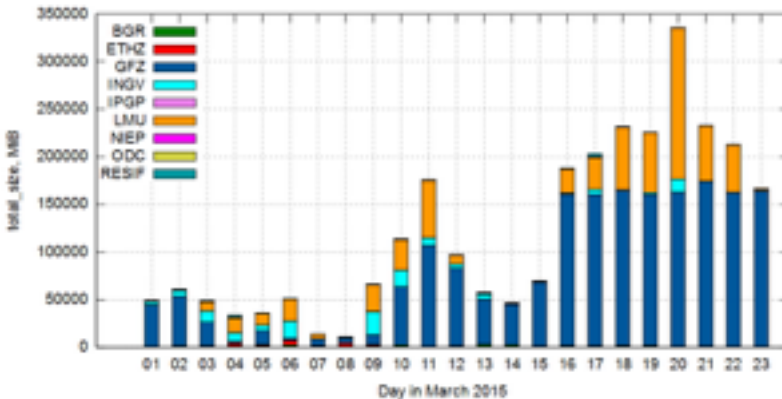
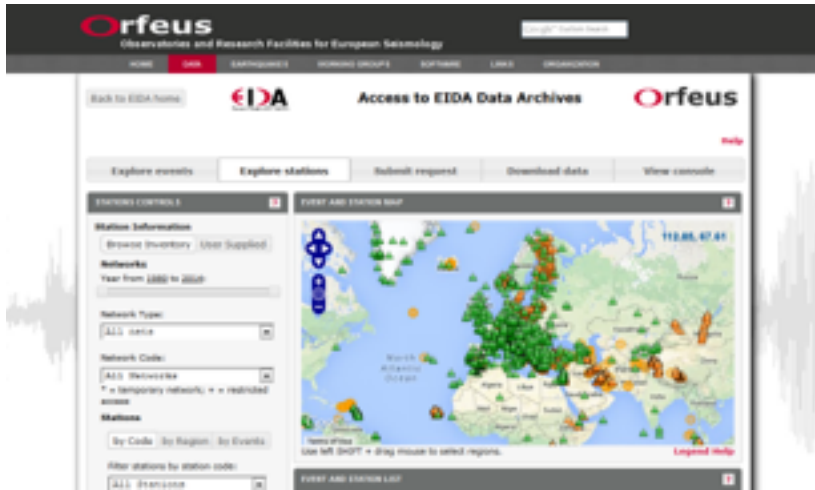
- **secure** and **long term** preservation
- data **curation**, distribution and access
- **acknowledgment** to data providers and **citation**












# EIDA mission





# EIDA some numbers

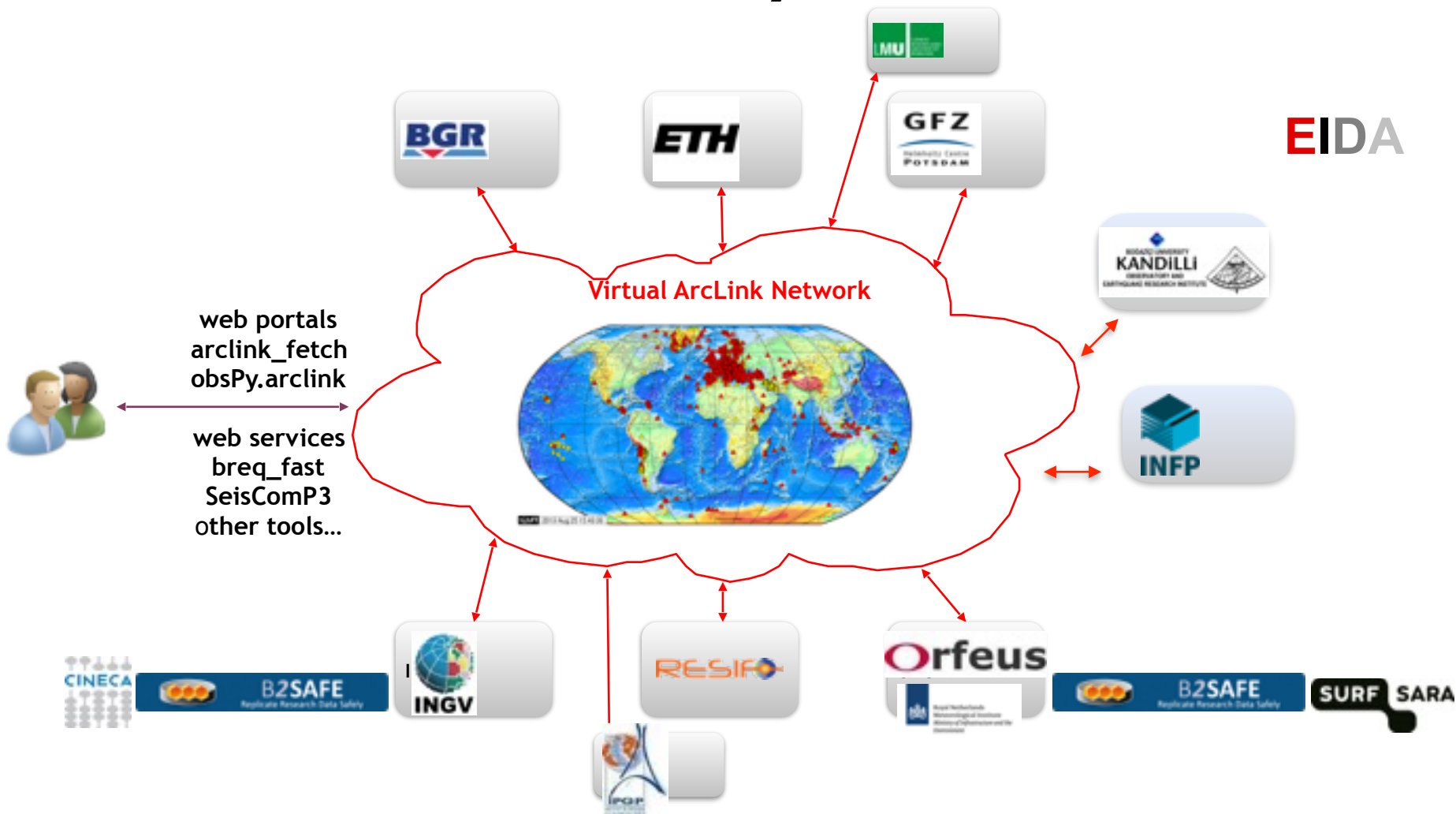


-  Federated archive for seismic waveforms
-  10 primary nodes, more to be added:
  -  Turkey added recently
  -  Greece in progress
-  ~ 5000+ stations
-  ~ 360+ TB total size
-  Persistent, safe storage
-  Data access services
  -  Easy access for scientists
  -  Up to ~15k users/year
  -  Multiple access methods

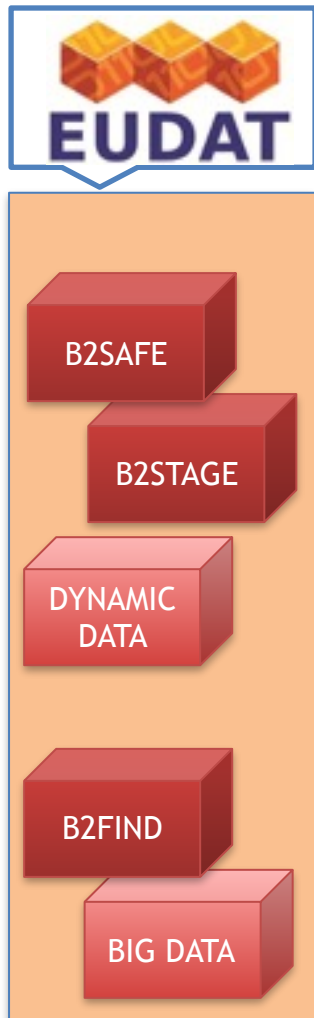
# EIDA today

EIDA

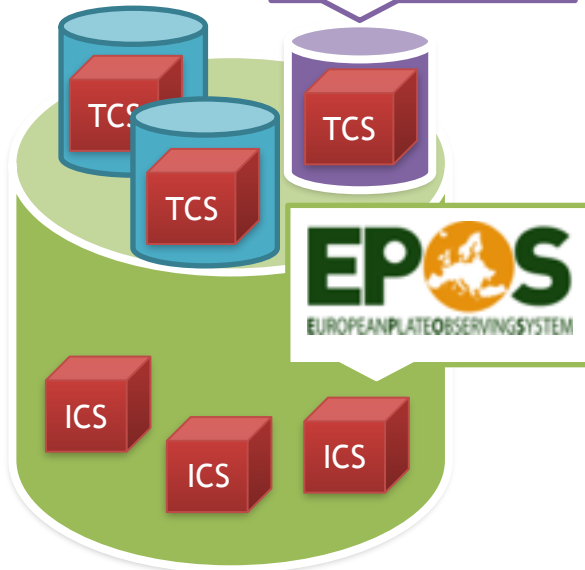
Users: Geoscientists etc...




# EUDAT – EPOS – EIDA relation








Any other specific community beyond EIDA within EPOS







## EUDAT:

-  Community independent
-  Generic data management services
-  Generic data discovery

## EPOS:

-  Solid Earth sciences
-  Community services integration
-  Organise and foster data standards and sharing and facility access
-  Cross-discipline data access (brokerage)
-  Multidisciplinary data-products, data modelling analysis and visualisation tools

## EIDA:

-  Seismology
-  Key element of the EPOS system
-  Fundamental service for higher level products
-  Seismic waveform data and metadata services



# EPOS-Seismology Uptake plan




KNMI-ODC , INGV and GFZ represent EPOS-S as partners of EUDAT2020

- Seismology has a very long tradition and widely recognised culture in data sharing
- Existing and widely accepted standards for data formats and exchange protocol
- Well-established governance
- Very large community within Earth sciences
- Shares products and methods with other disciplines
- Well-defined development plan

# EPOS-S Uptake Plan – main objectives

- Improve data **preservation**
- Improve data **discoverability**, **reuse** and access
- Guarantee failsafe and transparent access
- Achieve **identification**, **citation**, **traceability** of data and reproducibility of (scientific) products
- Harmonise data management policies across federation
- Facilitate data **movement** and analysis of large volumes of data
- Evolve EIDA services
- Subsequently extend the effort to other EPOS domains and services

# Starting point

-  **INGV** (CINECA) was partner on the 1<sup>st</sup> EUDAT project
-  **KNMI** (SURFsara) took part in one of the data pilots and implemented a preliminary B2Safe installation in that context
-  **GFZ** (KIT) joined EUDAT with the new EUDAT 2020 project

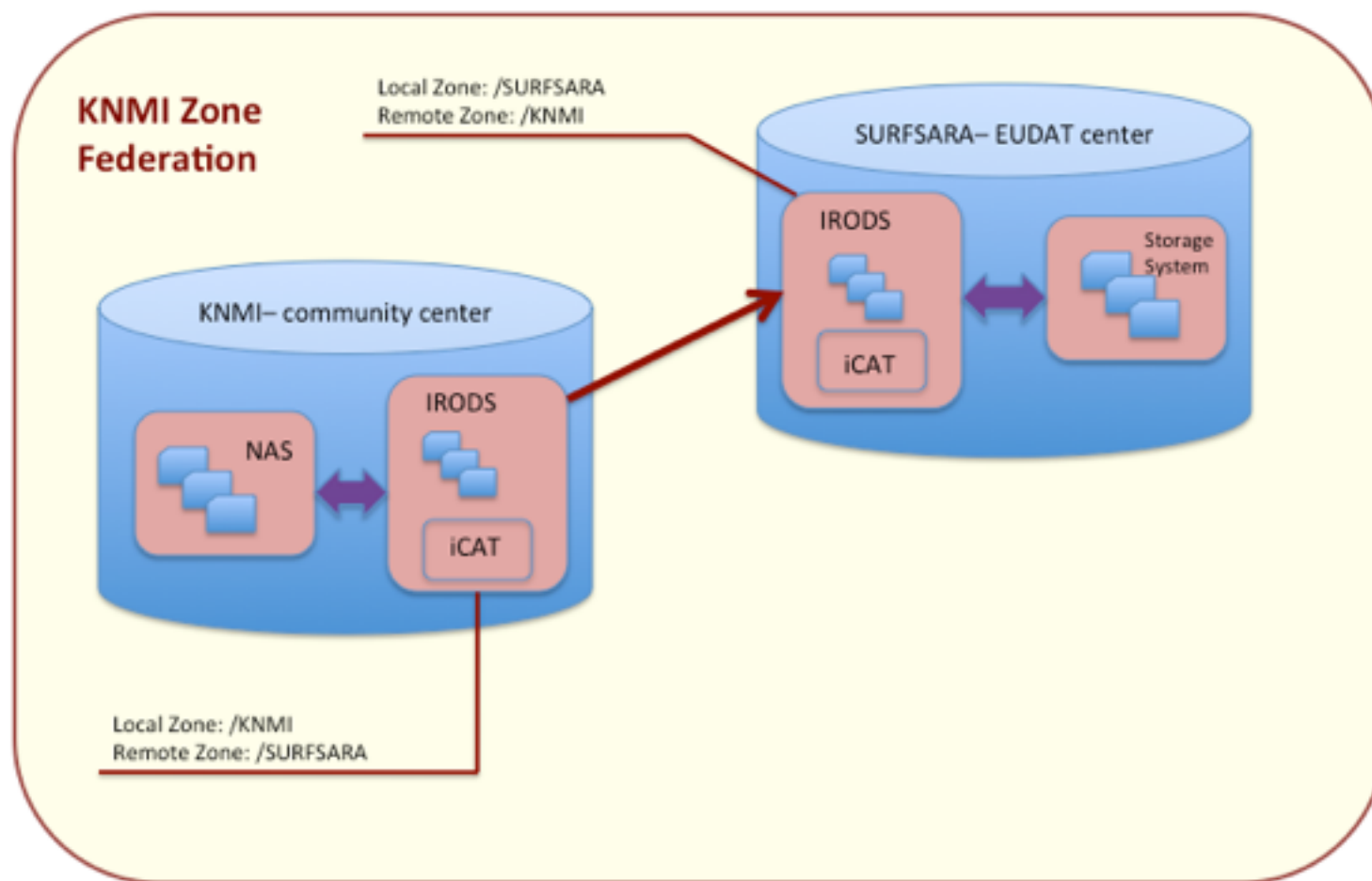


# Uptake plan current status

- Consolidating installation:  
Reinstalled, upgraded and reconfigured the B2Safe software stack, e.g. iRODS and B2Safe update
- Registering data object with their operational metadata to iCAT
- Adding community datacenter PID minting
- Fixing datacenter specific setups and issues
- Developing services on top of iRODS
- Testing the "*new federation*"
- Automating replication process

Main components:

- iRODS: micro services utilised to create customisable replication rules
- EPIC handle system: PID minting and management



# Workflow

- Register DO in iRODS (locally at KNMI)
- Generate PIDp at KNMI with own prefix
- Replicate DO to SURFsara
- Generate PIDr of the replica at SURFsara with “Repository of Record” populated (ROR)
- Update “Locations” in parent PIDp with replica information

In this way we are able to keep **crosslinks** between a certain dataset or DO and its replica(s) => useful for **further applications**

# PIDs - example

## Handle System®

Handle Values for: 11230/51a077d0-278b-11e4-be26-d89d6771dd88

Index	Type	Timestamp	Data
1	<a href="#">URL</a>	2014-08-19 10:26:45Z	<a href="irods://bhlsa08.knmi.nl:1247/ORFEUS/eudat/data/continuous/2014/001/AAK.BH1_00.II.2014.001">irods://bhlsa08.knmi.nl:1247/ORFEUS/eudat/data/continuous/2014/001/AAK.BH1_00.II.2014.001</a>
2	<b>CHECKSUM</b>	2014-08-19 10:26:45Z	3b1e53cc59c606439dac61ed02f24ef0
3	<a href="#">10320/LOC</a>	2014-08-19 12:46:45Z	<locations><location href="irods://bhlsa08.knmi.nl:1247/ORFEUS/eudat/data/continuous/2014/001/AAK.BH1_00.II.2014.001" id="0"/><location href="http://hdl.handle.net/11112/b36d2be4-279e-11e4-af1e-a0369f0b5f26" id="1"/></locations>
100	<a href="#">HS_ADMIN</a>	2014-08-19 10:26:45Z	handle=0.NA/11230; index=200; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]

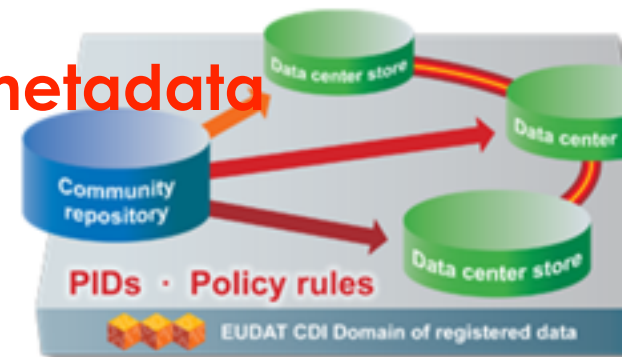
## Handle System®

Handle Values for: 11112/b36d2be4-279e-11e4-af1e-a0369f0b5f26

Index	Type	Timestamp	Data
1	<a href="#">URL</a>	2014-08-19 12:45:29Z	<a href="irods://irods1.storage.sara.nl:1247/vzSARA1/eudat/knmi/2014/001/AAK.BH1_00.II.2014.001">irods://irods1.storage.sara.nl:1247/vzSARA1/eudat/knmi/2014/001/AAK.BH1_00.II.2014.001</a>
2	<b>CHECKSUM</b>	2014-08-19 12:45:29Z	3b1e53cc59c606439dac61ed02f24ef0
3	<a href="#">10320/LOC</a>	2014-08-19 12:45:30Z	<locations><location id="0" href="irods://irods1.storage.sara.nl:1247/vzSARA1/eudat/knmi/2014/001/AAK.BH1_00.II.2014.001"/></locations>
4	<a href="#">EUDAT/ROR</a>	2014-08-19 12:45:31Z	<a href="http://hdl.handle.net/11230/51a077d0-278b-11e4-be26-d89d6771dd88">http://hdl.handle.net/11230/51a077d0-278b-11e4-be26-d89d6771dd88</a>
5	<a href="#">EUDAT/PPID</a>	2014-08-19 12:45:31Z	11230/51a077d0-278b-11e4-be26-d89d6771dd88
100	<a href="#">HS_ADMIN</a>	2014-08-19 12:45:29Z	handle=0.NA/11112; index=200; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]









# Open points

- PID assigned only to archived datasets. Datasets are assumed as frozen at a certain point in time. Lack of **update policies**
- We are currently working on a solution to embed automatic replication in our **data management procedures**
- Replication applied only to continuous waveforms but it could be extended also to other products
- PID granularity - currently PIDs are associated to daily files
- Replication mainly for **preservation** but we want to do more!
- PID not linked (yet) to domain specific **metadata**
- **Dynamic Data** issue not tackled yet





## - **beyond preservation**

- 
 Implementing the **B2Safe** service to ensure preservation of archived data
- 
 B2Safe to serve as base for further developments and activities
- 
 Allow **discovery** and reuse of the data ingested into B2Safe
- 
 Broaden current scope to cover more aspects of the Data Lifecycle adding new Data Management Policies (**DMP**)
- 
 Enable traceability of data **curation** steps
- 
 Synchronisation, verification and quality check of replicas, versioning
- 
 Integrate community services
- 
 And more...





- Harvesting existing community catalogs
- Integrated with B2SHARE metadata
- We would like to see a connection with B2SAFE via metadata
- Link to B2Note for user annotations and tagging
- Citation Use Case “packaging data collections for publication”



# EUDAT B2Find Data Citation use case

“A researcher browse the B2FIND catalog to discover datasets of interests. The researcher collects in his personal space the datasets matching his search criteria.

The researcher can bundle/group/ (re-)shuffle datasets and collections and annotate them.

Specific annotations could contain for instance the DOI of a scientific publication where the selected datasets have been used”

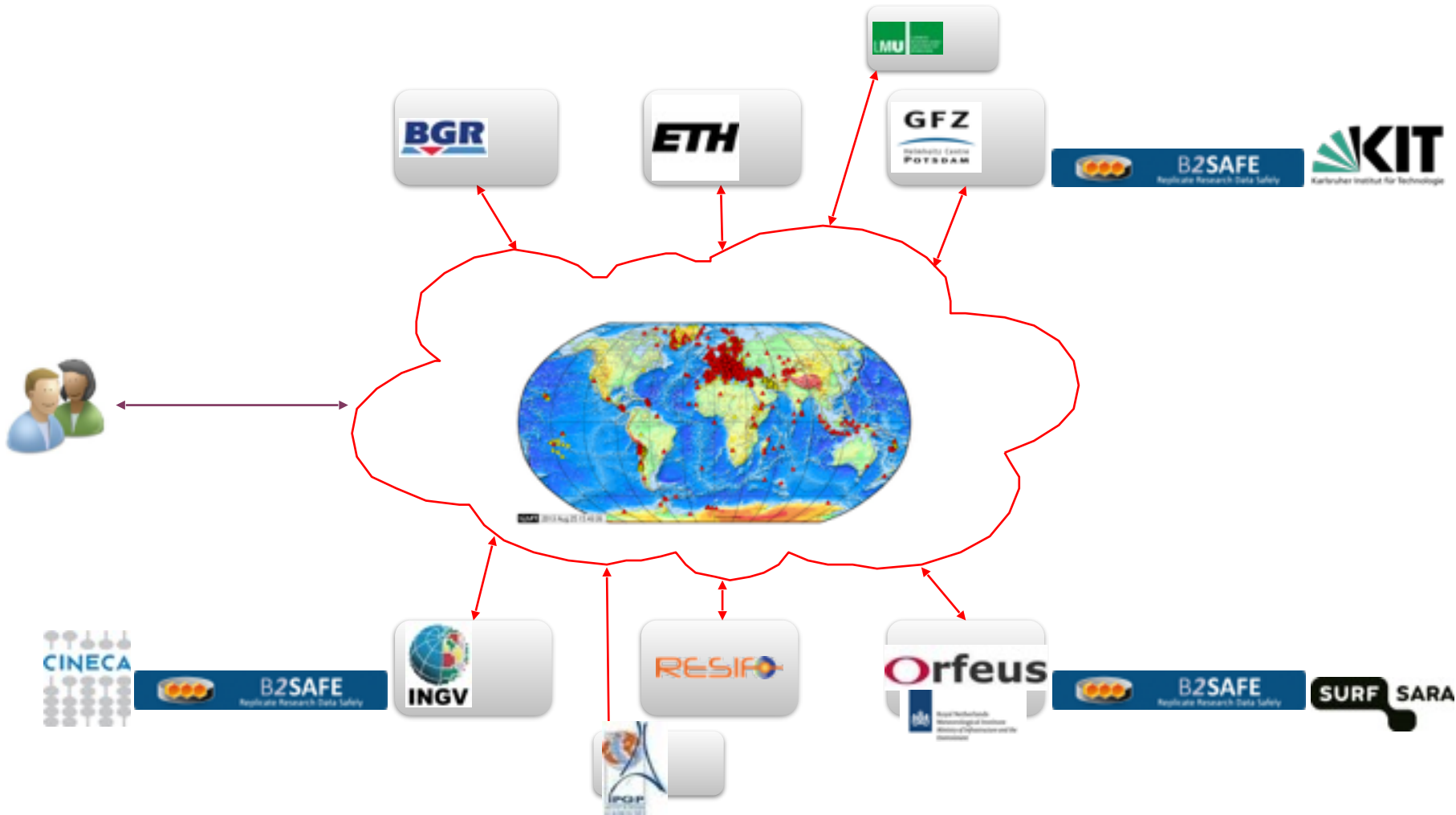
- B2FIND could help creating a link between scientific publications and data producers
- Promote dataset **citation** policies in scientific publications
- B2FIND could act as an enhanced aggregator and provide functionalities currently not present in domain specific catalogs
- Provide added value on top of community repositories. Eg: Find all the publications where dataset 'X' was used. Create statistics by dataset usage, by publisher etc...

# Other services

- B2ACCESS to enable access control and accounting  
Extend usage beyond CDI under discussion
- B2STAGE - to enable massive data processing reducing remote shipment and by means of PIDs
- Dynamic Data
- Semantic Services and Tools
- Workflows, computation, products generation via Generic Execution Framework
- Provenance
- Foster scientific collaboration and exchange of experiments and results via B2SHARE and B2DROP

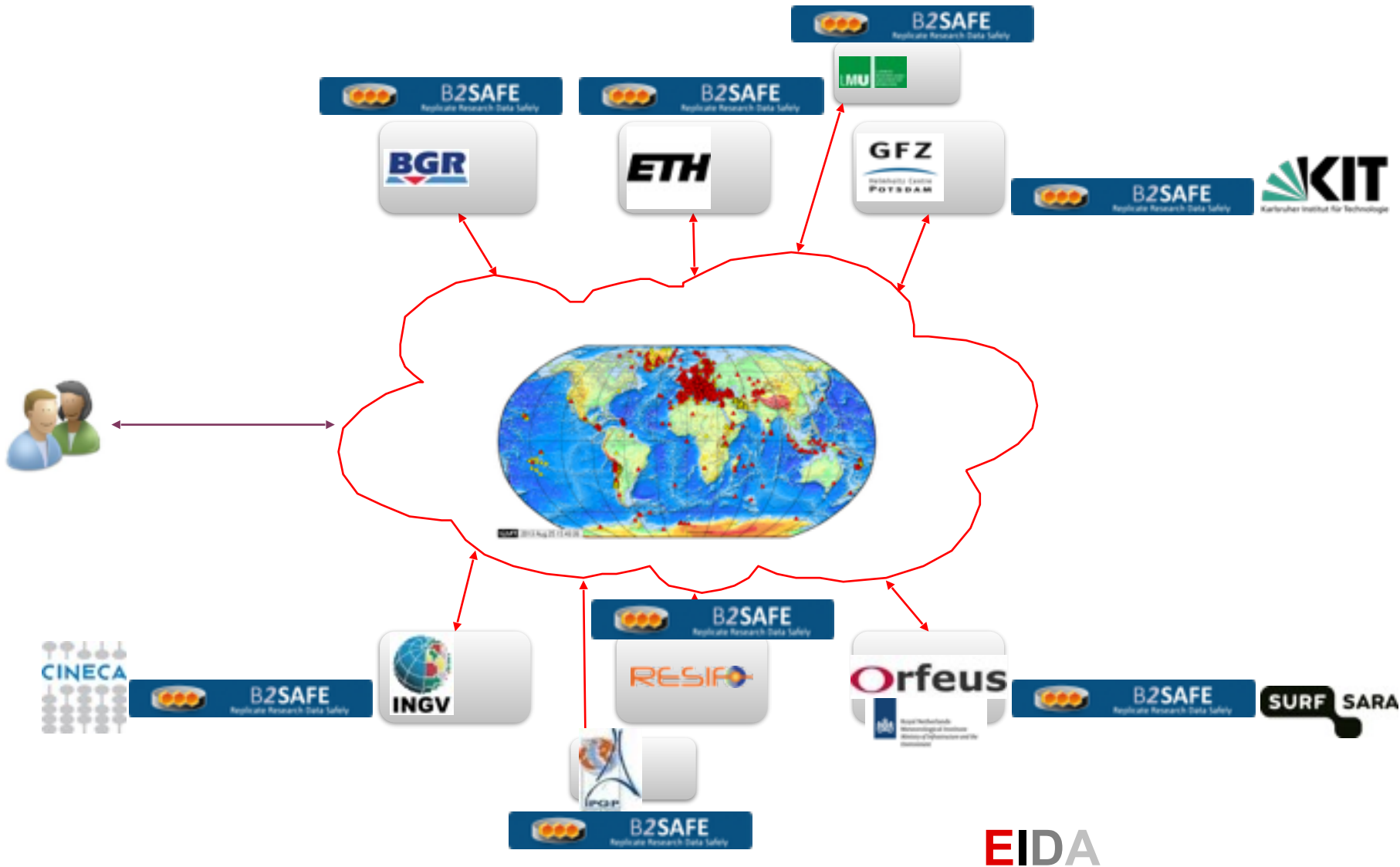
# EIDA in EUDAT2020

Users: Geoscientists etc...

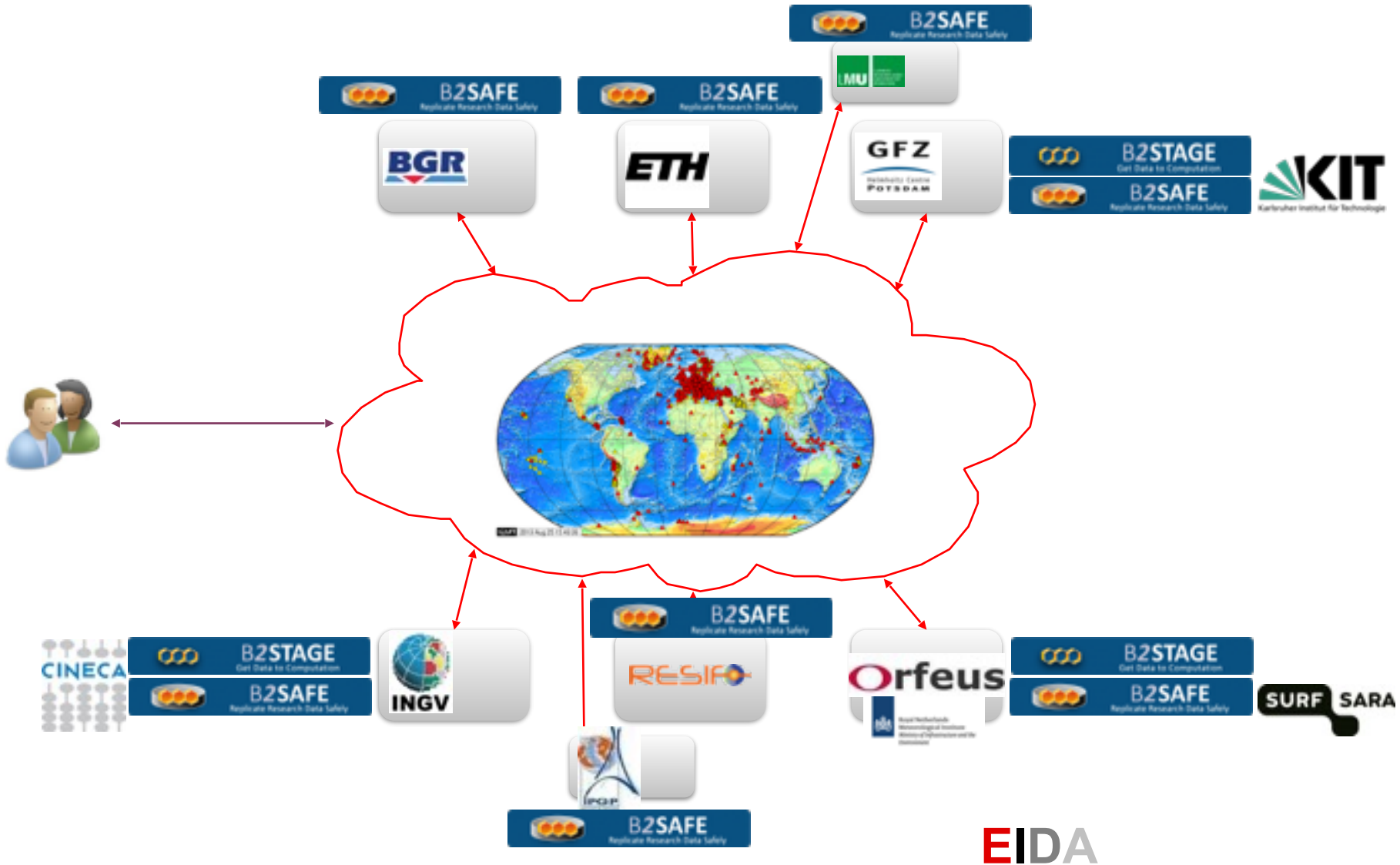


EIDA

Users: Geoscientists etc...

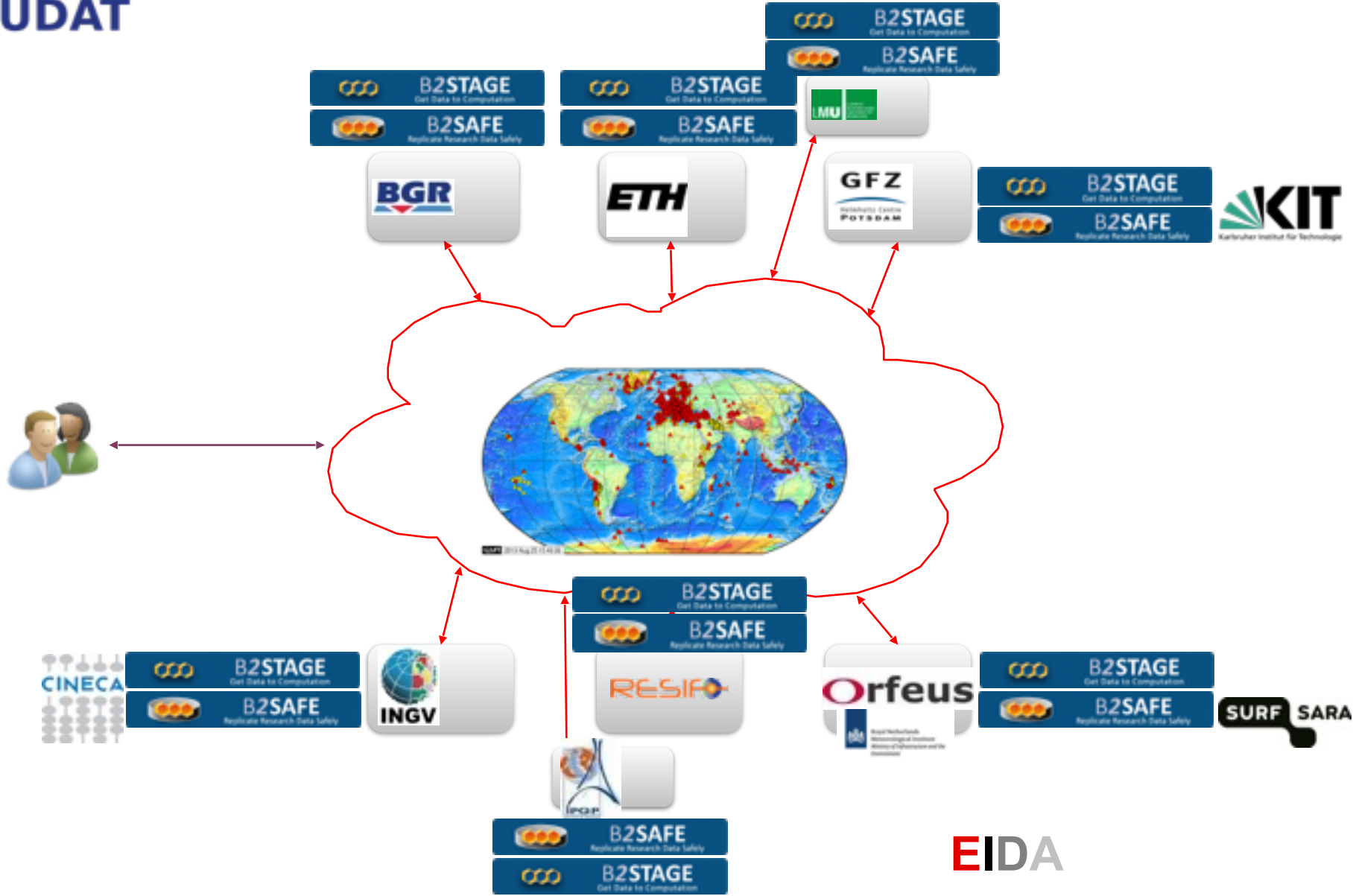


Users: Geoscientists etc...

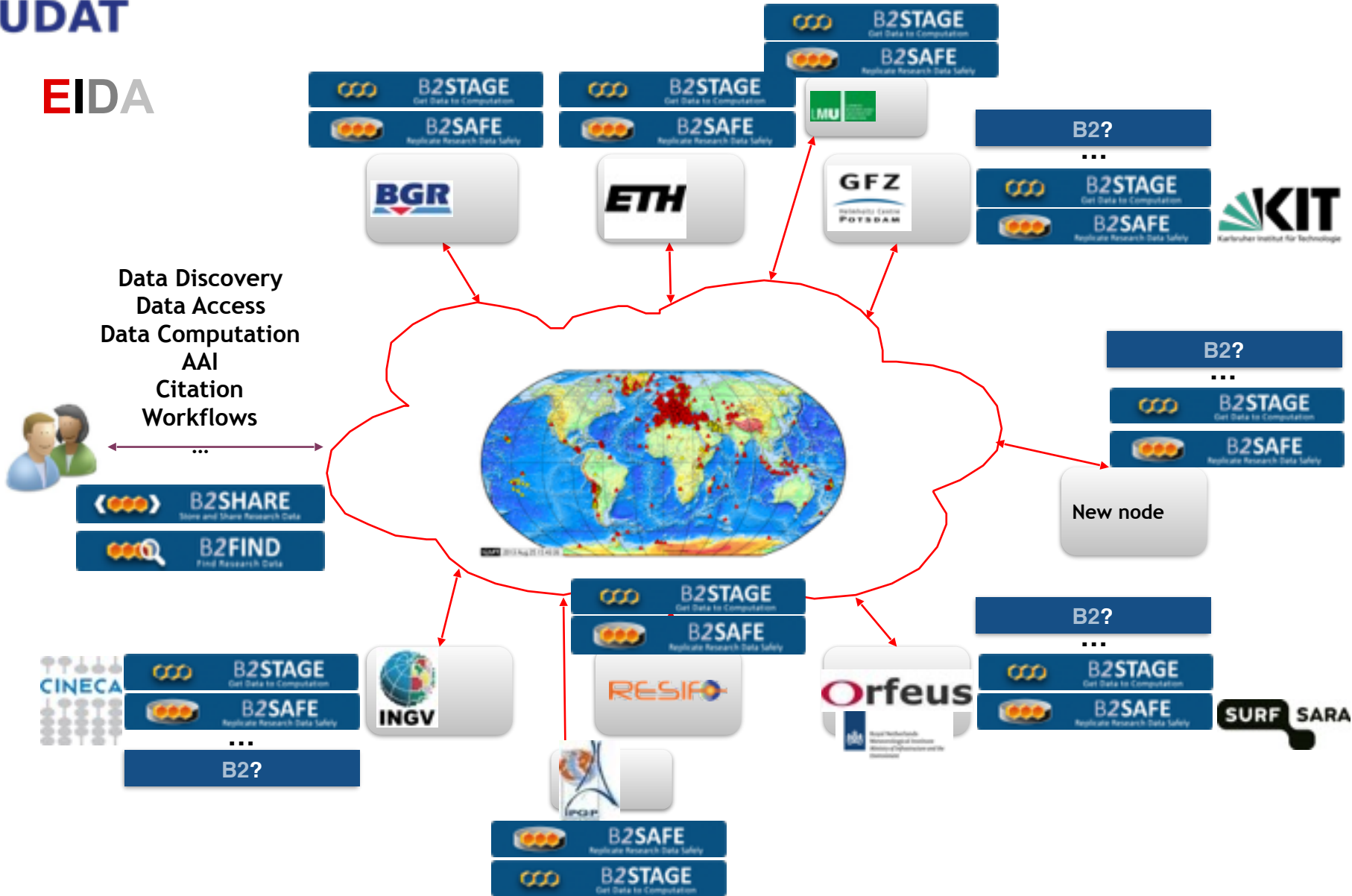




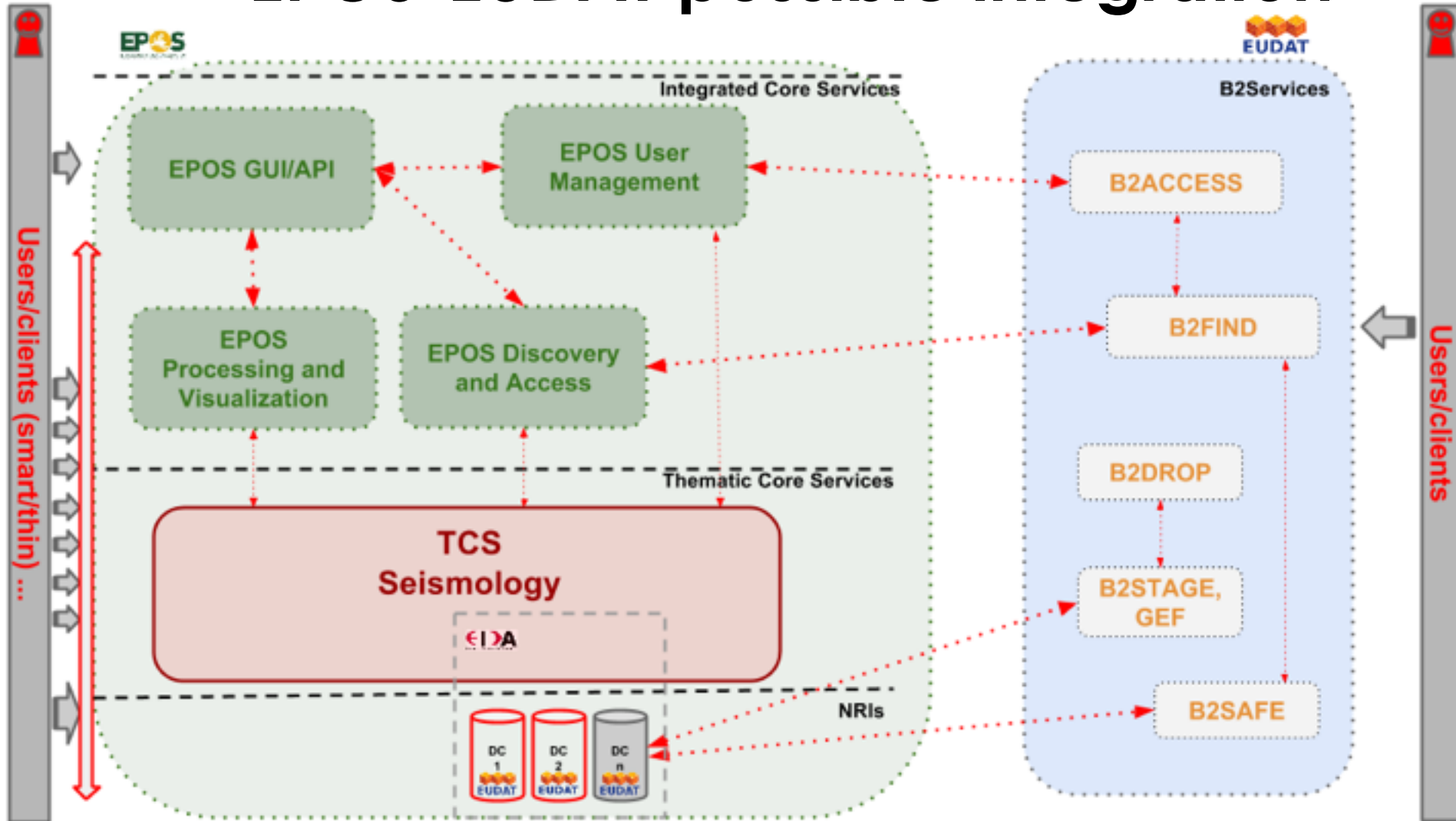
Users: Geoscientists etc...

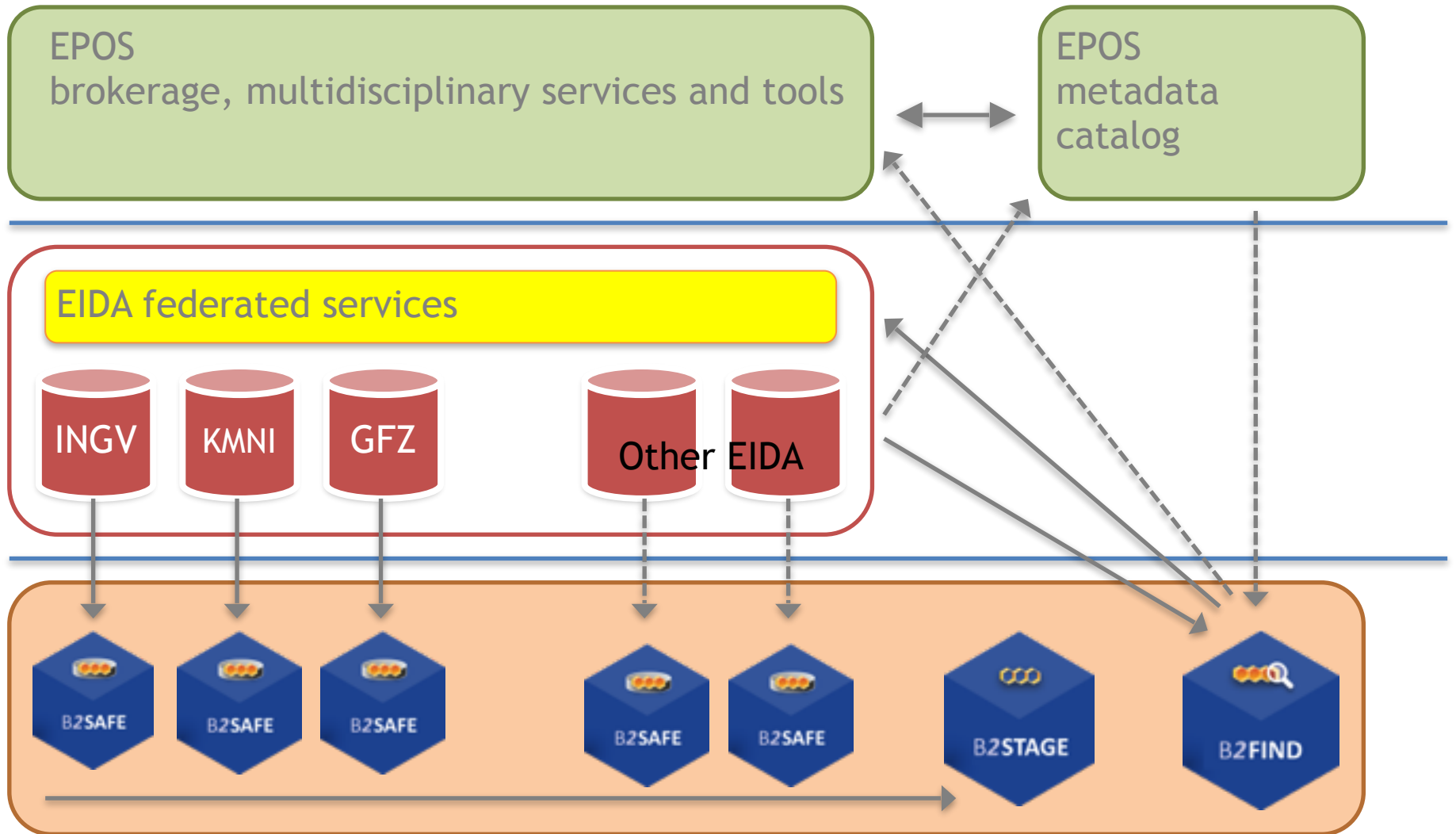


EIDA



# EPOS-EUDAT possible integration







# Seismic Streams

**Seismic Data Streams** are stored in heterogeneous ways, usually according to a well-known and widespread data format namely (FDSN) Standard for the Exchange of Earthquake Data or **SEED**

SEED (or MSEED in his compact version) is a **community standard** used to store and exchange Seismic Time Series (**waveforms**)

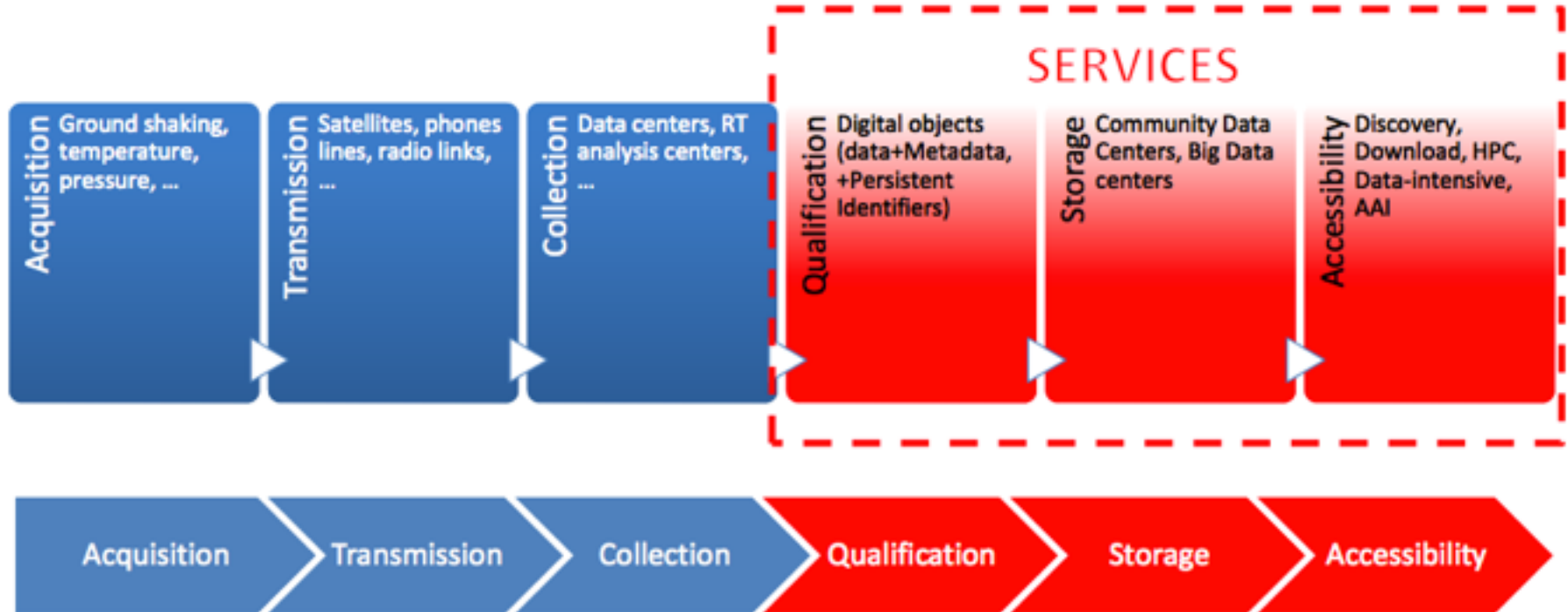
It is a **compressed** format based on STEIM compression algorithm: it holds differences between differences between 2 adjacent samples

MSEED organises a continuous seismic stream into chunks of fixed length (depending on the sample rate) called **Records**

A record consist of a **fixed header**, containing characteristics and metadata, and of a **payload** with **digital counts**

The **physical units** can be derived combining counts with the **instrument response** usually stored in a metadata structure according to another community standard schema: FDSN **stationXML**

# Timeline

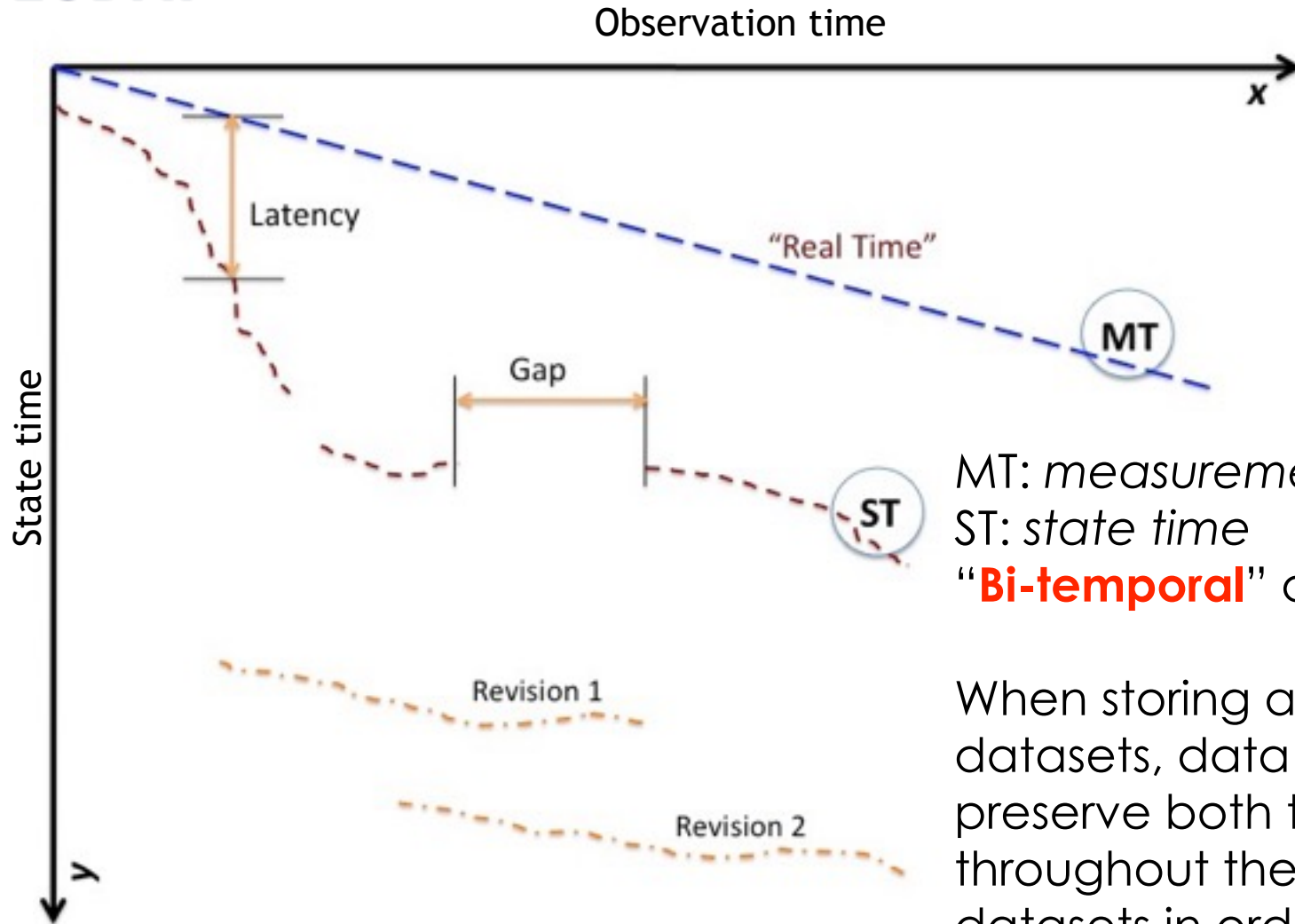




# Issues to be considered

- Data streams received not necessarily in real time and in some cases not even in sequence
- Data streams can contain gaps which are usually filled in a later stage
- There might be changes in the instrumentations (re-calibrations, manual corrections,...) affecting the data stream
- Changes might concern only certain channels

# Temporal challenge



MT: *measurement time*  
 ST: *state time*  
 “**Bi-temporal**” data.

When storing and handling the datasets, data centers must preserve both these times throughout the lifecycle of the datasets in order to reconstruct their history

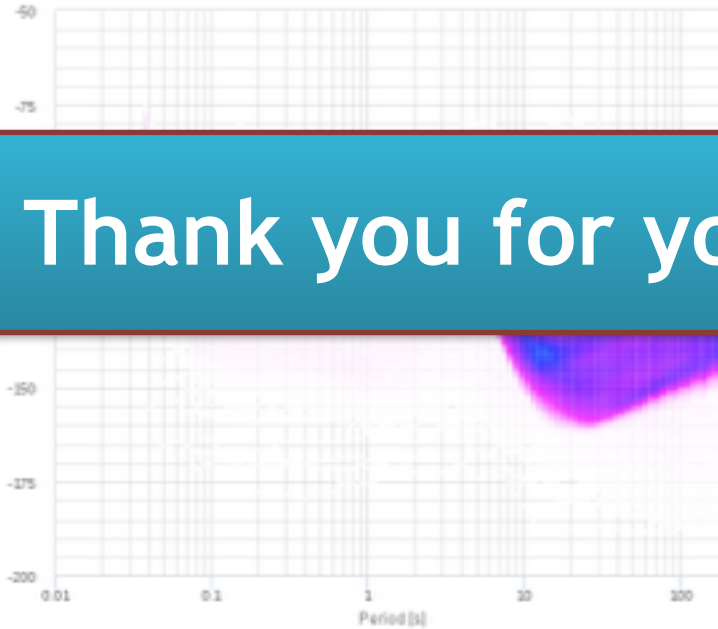
# Dynamic Data issue

- Datasets may be used for analysis or products generation even when partial and **incomplete**
- Challenging choice of clever strategy for PID assignment and management
- An important requirement is to keep the “history” or provenance for reproducibility and/or propagate later **refinements**
- Example of products include:
  - Power Spectral Density - 2D arrays with variation of instrument metadata(revisions) and spectral content (based on raw data)
  - Strong Motion - composite products including: raw and processed waveforms, site characteristics, instrument metadata, event information, more parameters
  - Shake maps - near real time 2D maps of ground motion and shaking intensity following significant earthquakes - based on SM processing
  - Cross-correlations

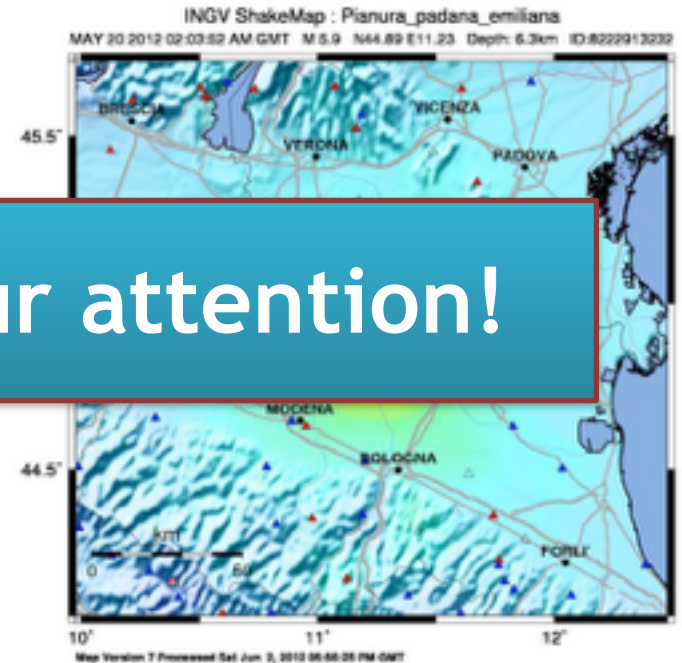
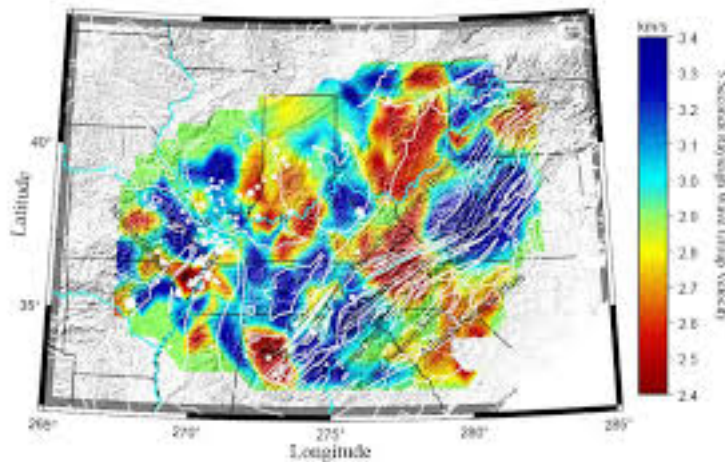
# Conclusions

- EUDAT provides generic and solid **building blocks** which can improve and boost the RIs architecture design
- Close collaboration with research communities ensures solutions targeted to specific and concrete requirements
- Wide platform to **share** knowledge and cross link heterogeneous scientific backgrounds
- Wide adoption of open source and standards coupled to strong support fosters dissemination and facilitates uptake
- EUDAT and EPOS well-established and promising mutual collaboration
- It is crucial to ensure sustainability!

PPSD of 1000 PSDs  
2011-01-01 to 2016-01-07



• Probabilistic Power Spectral Densities



PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
PERCEIVED SHAKING	none	none	none	Very light	Light	Moderate	Mod. Heavy	Heavy	Very heavy
PEAK ACC (mg)	<0.1	0.5	2.4	6.7	19	24	44	83	>155
PEAK VEL (mm/s)	<0.07	0.4	1.9	5.5	11	22	42	80	>150
ROTATIONAL VELOCITY	I	II-III	IV	V	VI	VII	VIII	IX	X+

Thank you for your attention!