Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and the Environment

B2Safe @ ODC and future EIDA/ EPOS-S plans within EUDAT2020

Luca Trani and the EIDA Team
Acknowledgements to SURFsara and the B2SAFE team

EIDA Orfeus
European Integrated Data Archives

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

3rd EUDAT Conference, Amsterdam, The Netherlands, 24-25 September 2014

Outline

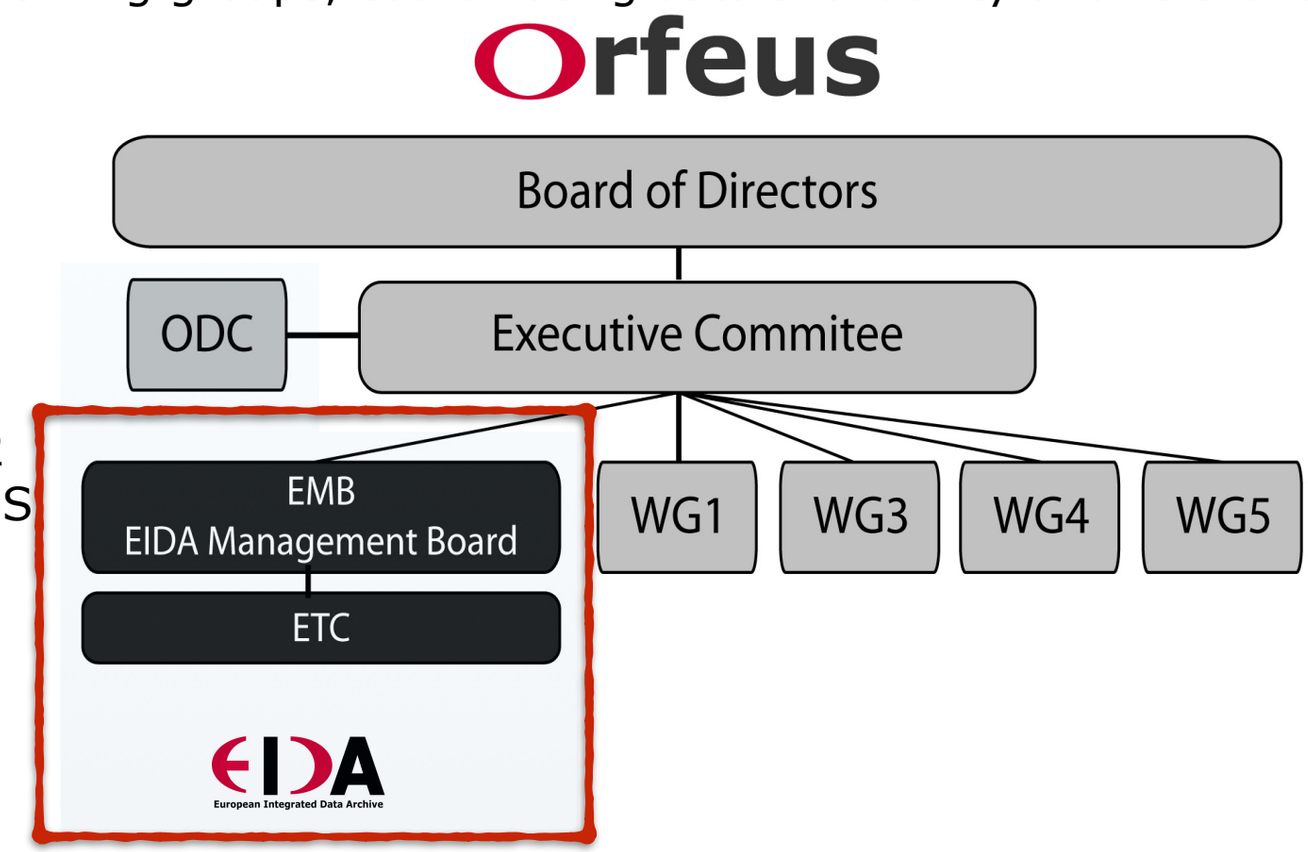
- ORFEUS
- European Integrated Data Archives concept
- ORFEUS Data Center
- Why B2SAFE and description of the implemented solution
- Current limitations
- EIDA/EPOS-S future plans within EUDAT2020

ORFEUS

ORFEUS (Observatories and Research Facilities for European Seismology), is the non-profit foundation that aims at co-ordinating and promoting digital, broadband (BB) seismology in the European-Mediterranean area.

It was founded in 1987 by corporate founders from 13 European countries forming the board of directors.

Its activities are distributed between the ORFEUS Data Centre (ODC), gathering, archiving and providing waveform data, and four working groups, coordinating data availability and relevant developments.



EIDA since Nov 2012
officially under ORFEUS

EIDA GOALs

- **safe, persistent** archival and dissemination of **high quality** seismic waveforms and products via **distributed archives**
- easy access for scientists - support multiple access methods, standards and tools
- open, where possible, but also closed / restricted access
- provide best datasets available

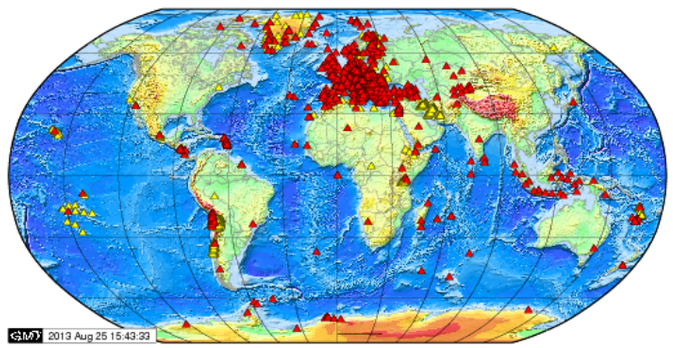
EIDA today

EIDA

Users: Geoscientists etc…

web portals
arclink_fetch
obsPy.arclink

web services
breq_fast
SeisComP3
other tools…



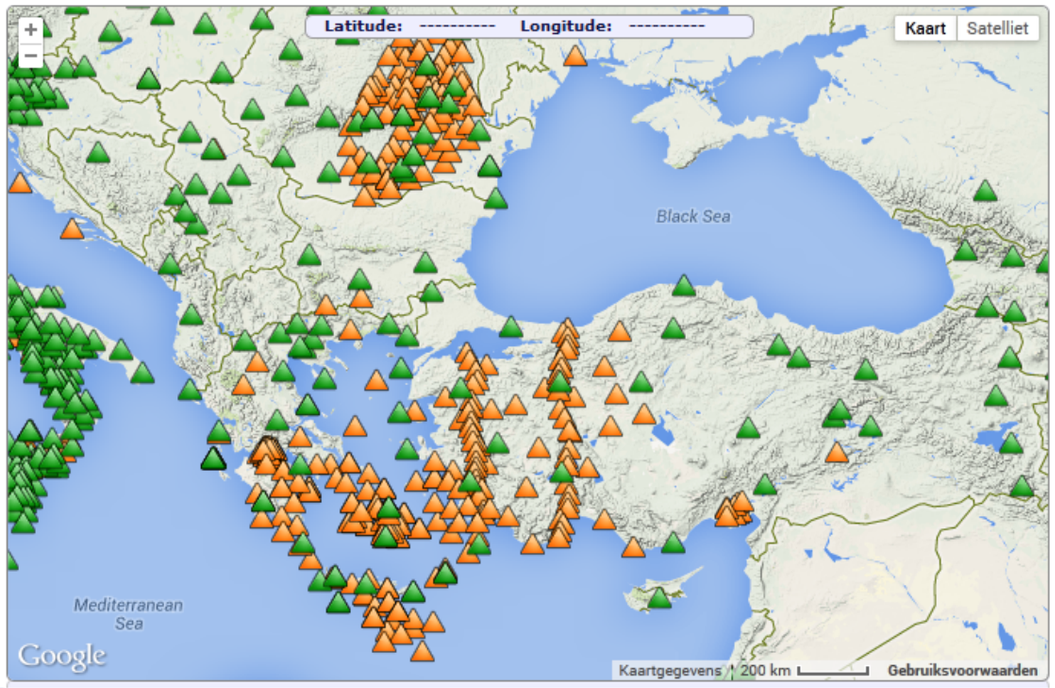
Virtual ArcLink Network

BGR
ETH
LMU
GFZ Helmholtz Centre Potsdam
INGV
RESIF
IPGP
Orfeus
Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment
CINECA
B2SAFE Replicate Research Data Safely
B2SAFE Replicate Research Data Safely
SURF SARA

EIDA in numbers

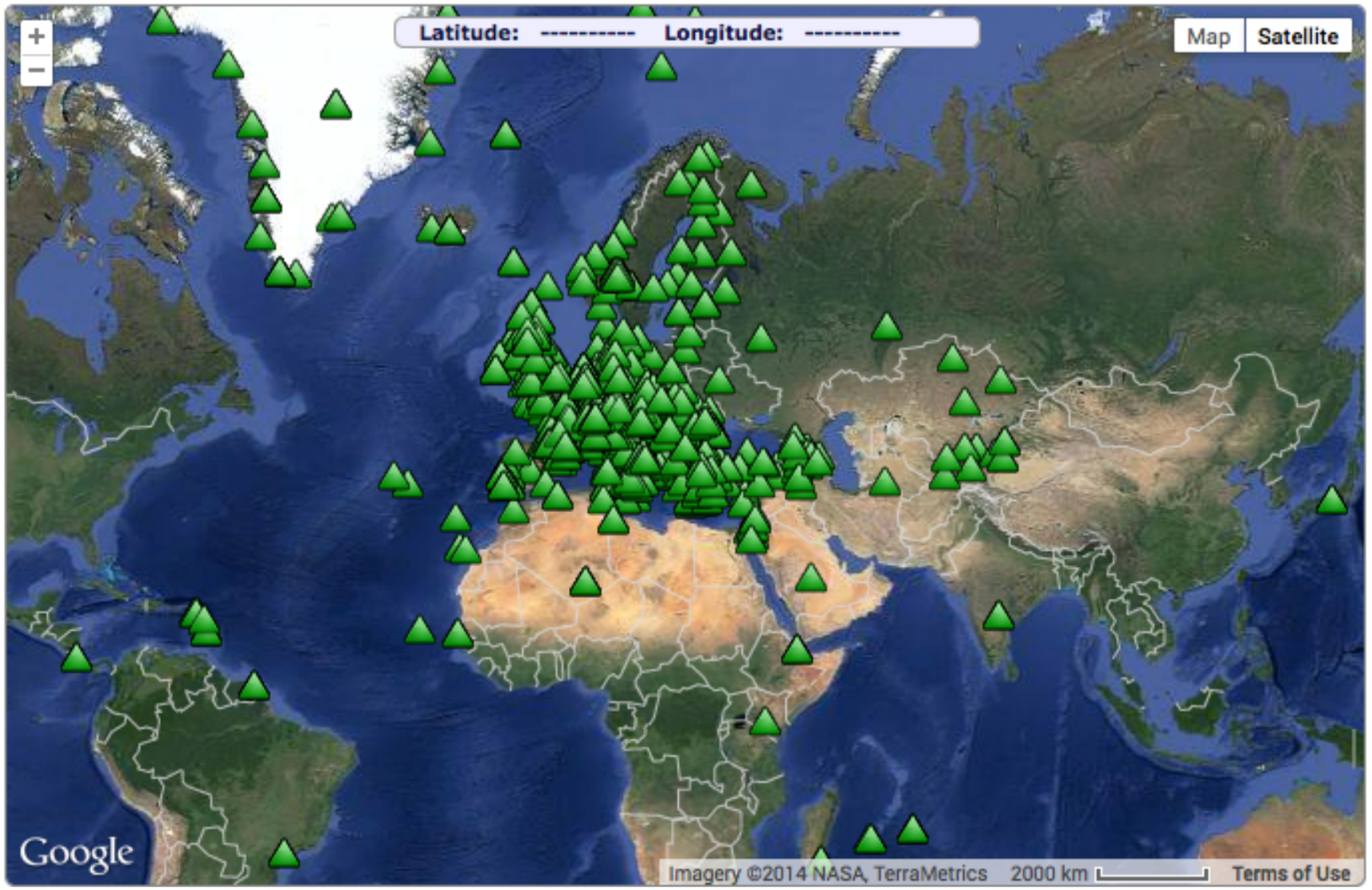
EIDA currently provides uniform, unrestricted and rapid access to about ~***400*** TB of data

- ***80*** permanent networks
- ***58*** temporary networks
- ***4500*** open access stations
- ***1600*** real-time permanent stations
- **25-50** GB downloads per day



Permanent station
Temporary station
EIDA node

ORFEUS Data Center



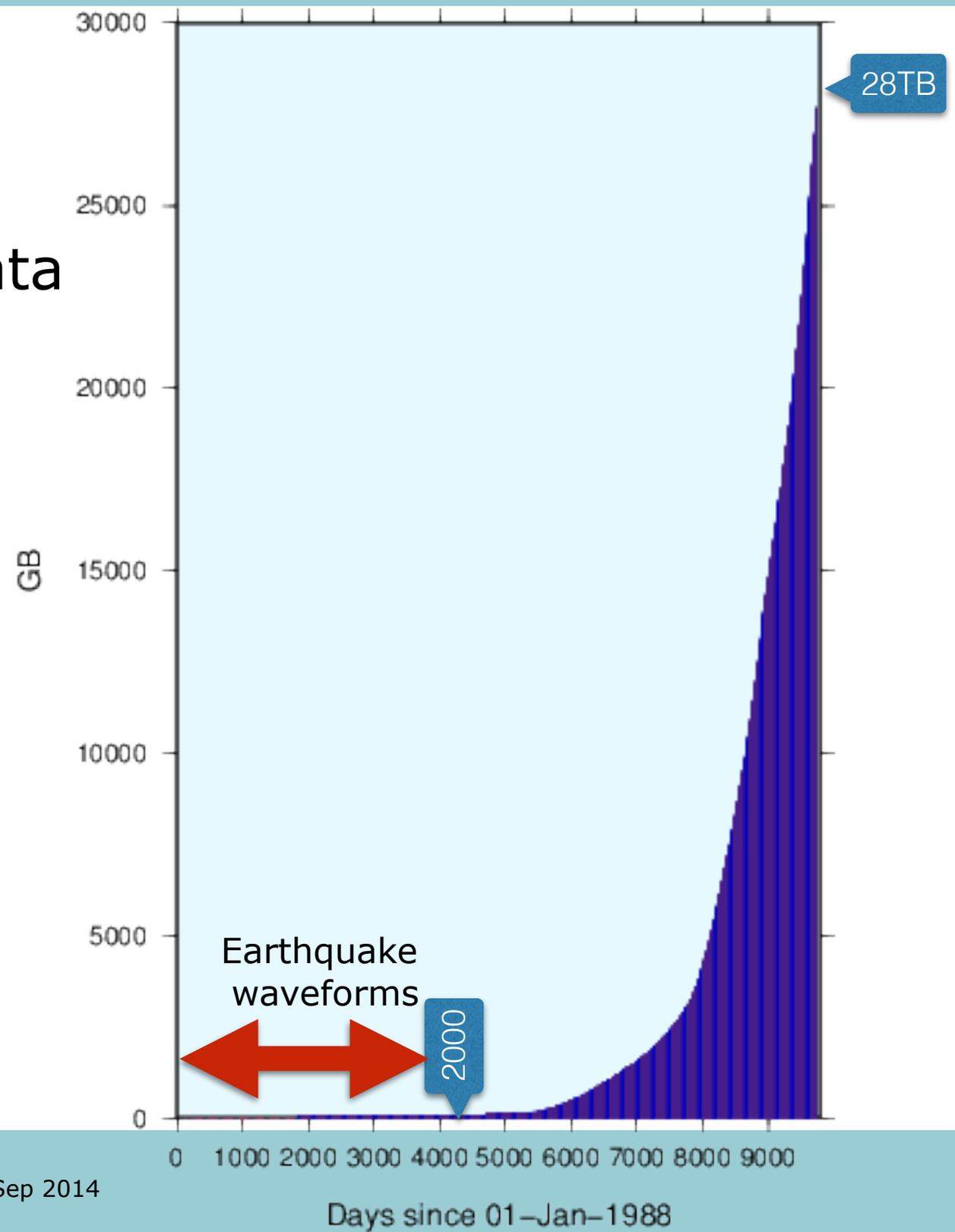
The Orfeus Data Center (ODC) acts as regional data centre within the International Federation of Digital Seismograph Networks (FDSN) and is hosted by the seismological division of the KNMI in The Netherlands

The VEBSN (Virtual European Broadband Seismograph Station Network) is the pool of broadband seismographs from which data is shared in real-time with ORFEUS Data Center

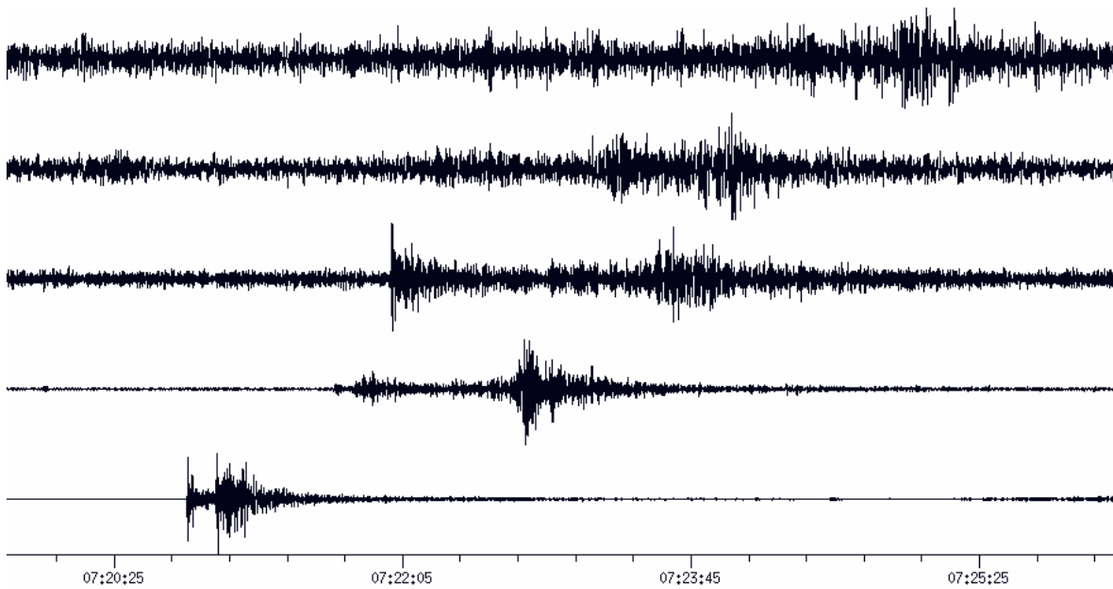
The centralised archive contains waveform data from 1988 to present, from about 60 networks

ODC: waveform data

Continuous waveforms data growth from 1988



Seismic continuous waveform



Seismic stations continuously record ground motion and send their data in real time to data centers which collect them

Each station is equipped with different sensors: typically targeting different spatial and frequency components

Data are usually stored as binary files encoded in a standard format called SEED (Standard for the Exchange of Earthquake Data) or better MSEED (its compact version)

A common convention is to arrange data in daily files divided per network,station,channel, location code

E.g.: SABA.HHE.NA.2014.001, SABA.HHN.NA.2014.001, SABA.HHZ.NA.2014.001

Why B2SAFE

Orfeus mission:

"Provide safe, persistent and long term archival and preservation of high quality waveform data"

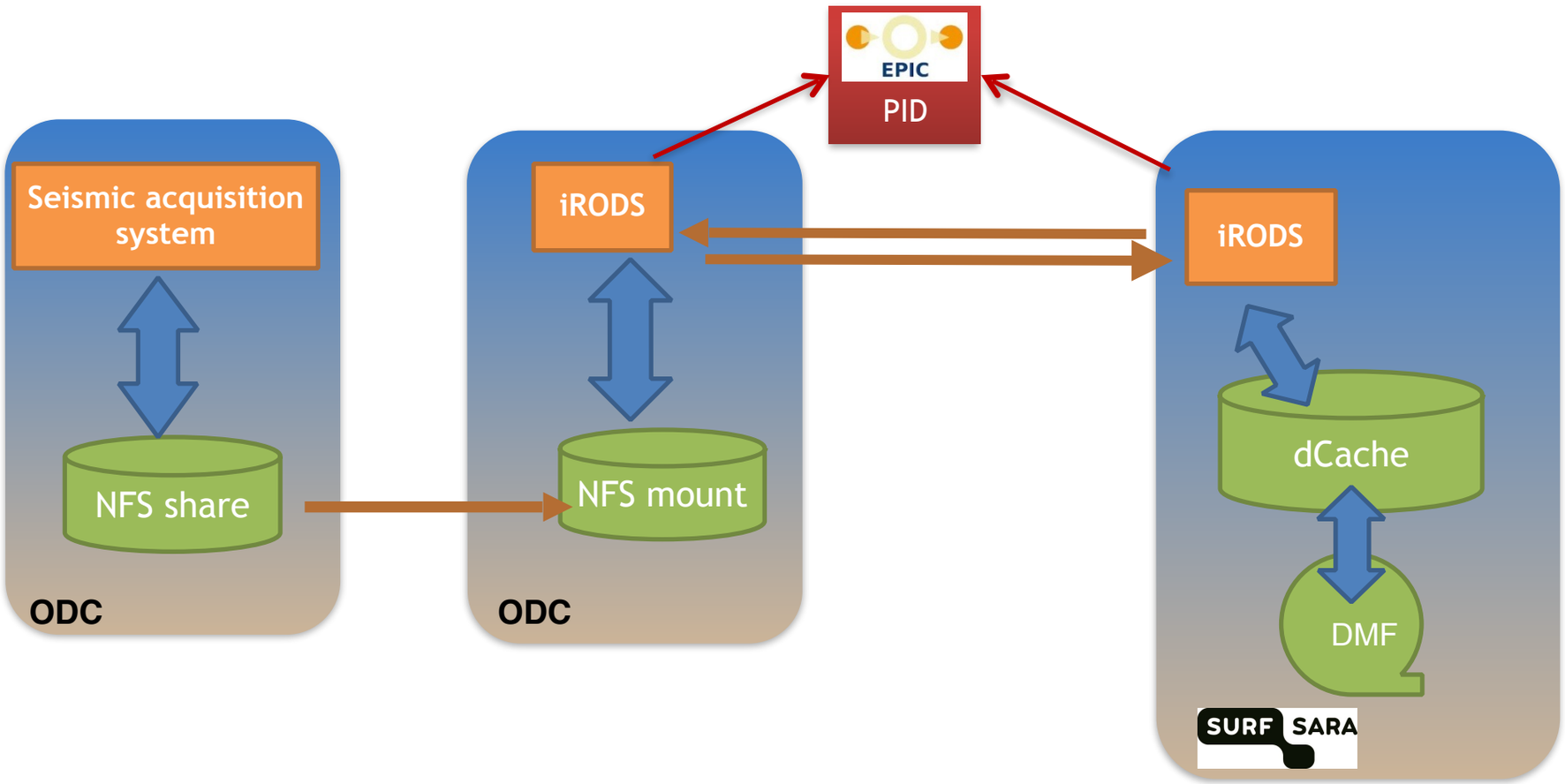
It matches perfectly with the features offered by B2SAFE

Moreover it allows full control on the choice of data replication facility vs generic cloud service => very important for computation

Possibility to implement customised rules and data workflows

Solution: overview

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



Thanks to Robert Verkerk

Solution: workflow

1) Register dataset in iRODS (locally at KNMI): ireg…

2) Generate PID at KNMI

3) Replicate dataset to SURFsara

4) Generate PID of the replica at SURFsara

5) Update parent PID with replica information at KNMI

Solution: PIDs

Handle System®

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

Index	Type	Timestamp	Data
1	**URL**	2014-08-19 10:26:45Z	irods://bhlsa08.knmi.nl:1247/ORFEUS/eudat/data/continuous/2014/001/AAK.BH1_00.II.2014.001
2	**CHECKSUM**	2014-08-19 10:26:45Z	3b1c53cc59c606439dac61ed02f24ef0
3	**10320/LOC**	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	**HS_ADMIN**	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	**URL**	2014-08-19 12:45:29Z	irods://irods1.storage.sara.nl:1247/vzSARA1/eudat/knmi/2014/001/AAK.BH1_00.II.2014.001
2	**CHECKSUM**	2014-08-19 12:45:29Z	3b1c53cc59c606439dac61ed02f24ef0
3	**10320/LOC**	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	**EUDAT/ROR**	2014-08-19 12:45:31Z	http://hdl.handle.net/11230/51a077d0-278b-11e4-be26-d89d6771dd88
5	**EUDAT/PPID**	2014-08-19 12:45:31Z	11230/51a077d0-278b-11e4-be26-d89d6771dd88
100	**HS_ADMIN**	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]

Solution: some numbers

- Replication of the ODC archive from 1988 to Dec 2013 completed ~ 6.6 M files

- Registration of the first 6 months of 2014 into iRODS at KNMI ~800.000 files

- Registration rate in iRODS ~100.000 files per day

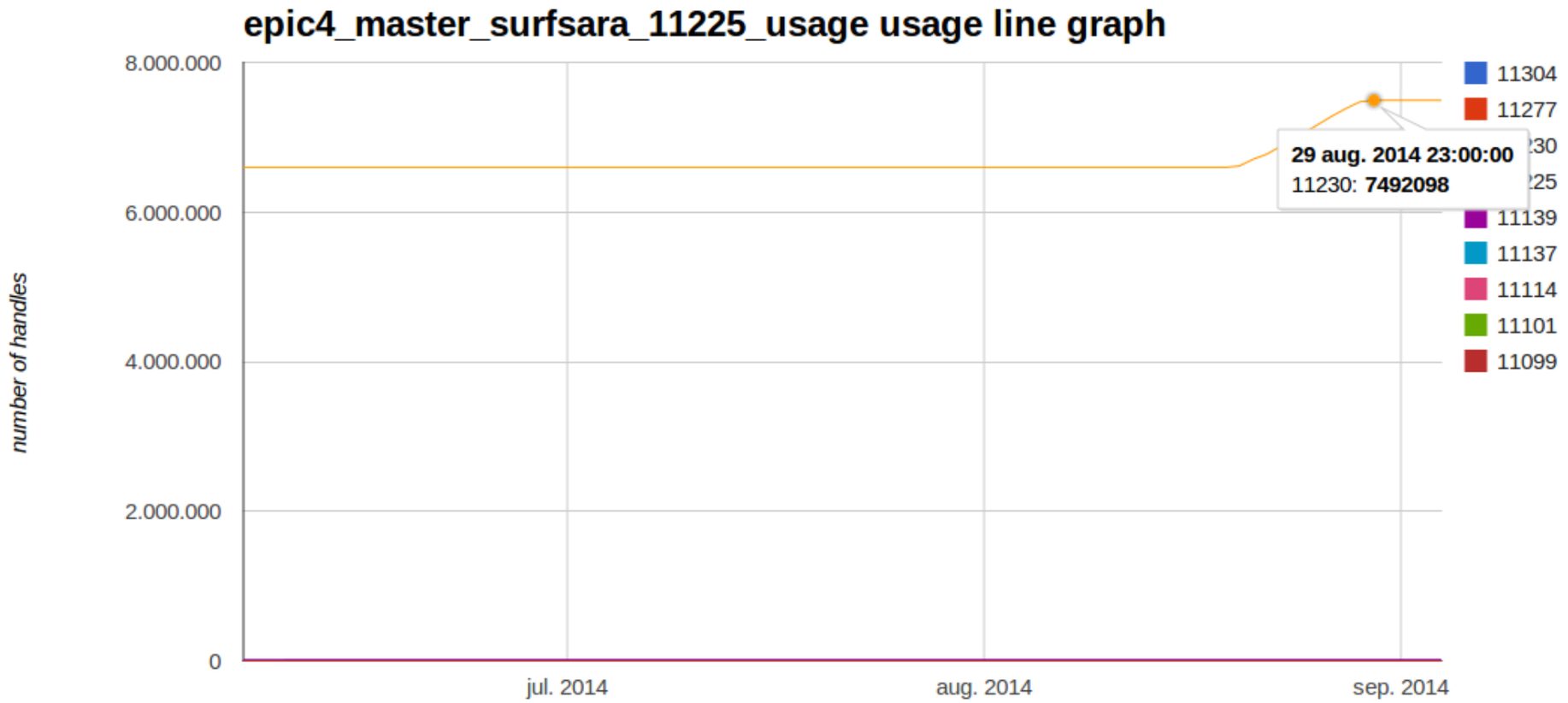
- Replication rate ~ 80.000 files per day
  (Typical file size range: from a few KB up to 20-30MB)

Planned actions:

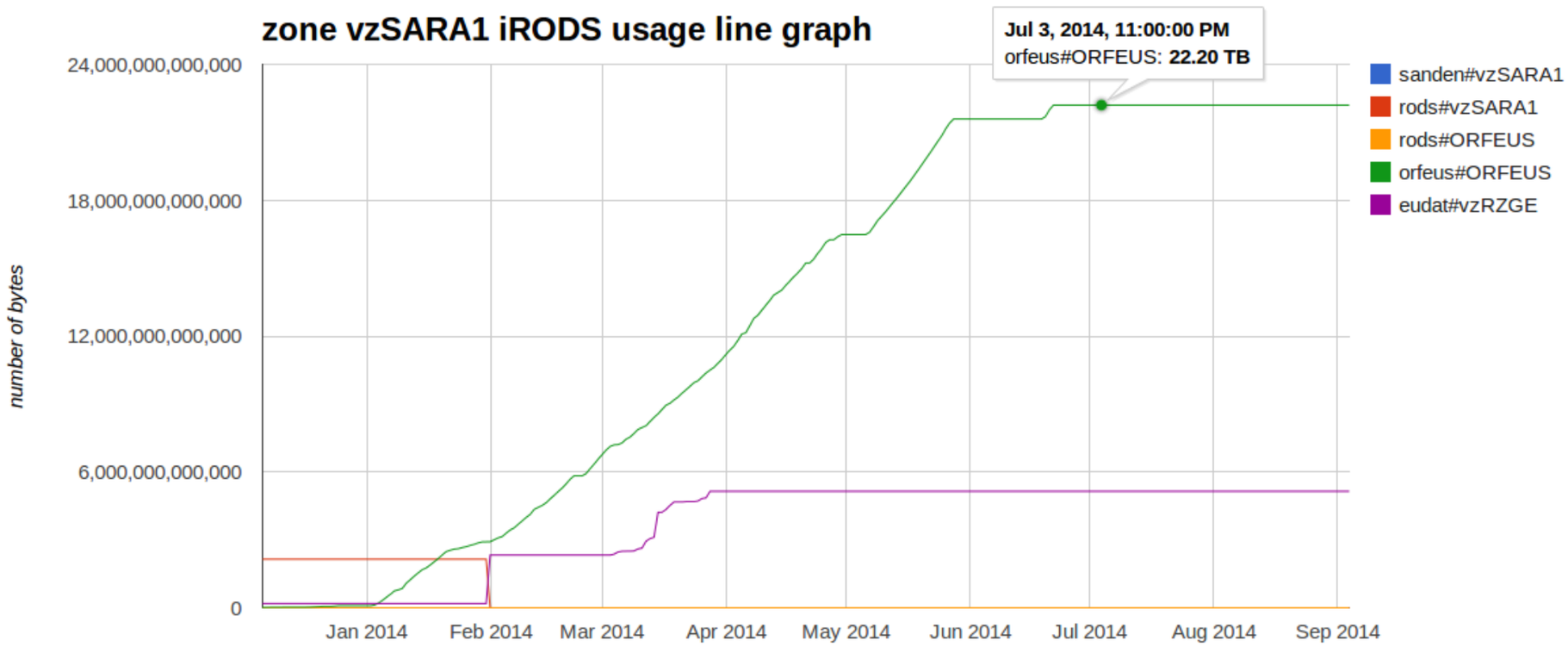
- Finalise replication up to current day

- Set up an automatic procedure to register and replicate new files

EPIC usage in handles



iRODS usage in bytes



Solution: current limitations

- PID assigned only to archived datasets: no real time yet

Datasets are assumed as frozen at a certain point in time

- Replication applied to continuous waveforms but it could be extended also to other products. E.g.: stations, qc, …

- PID granularity

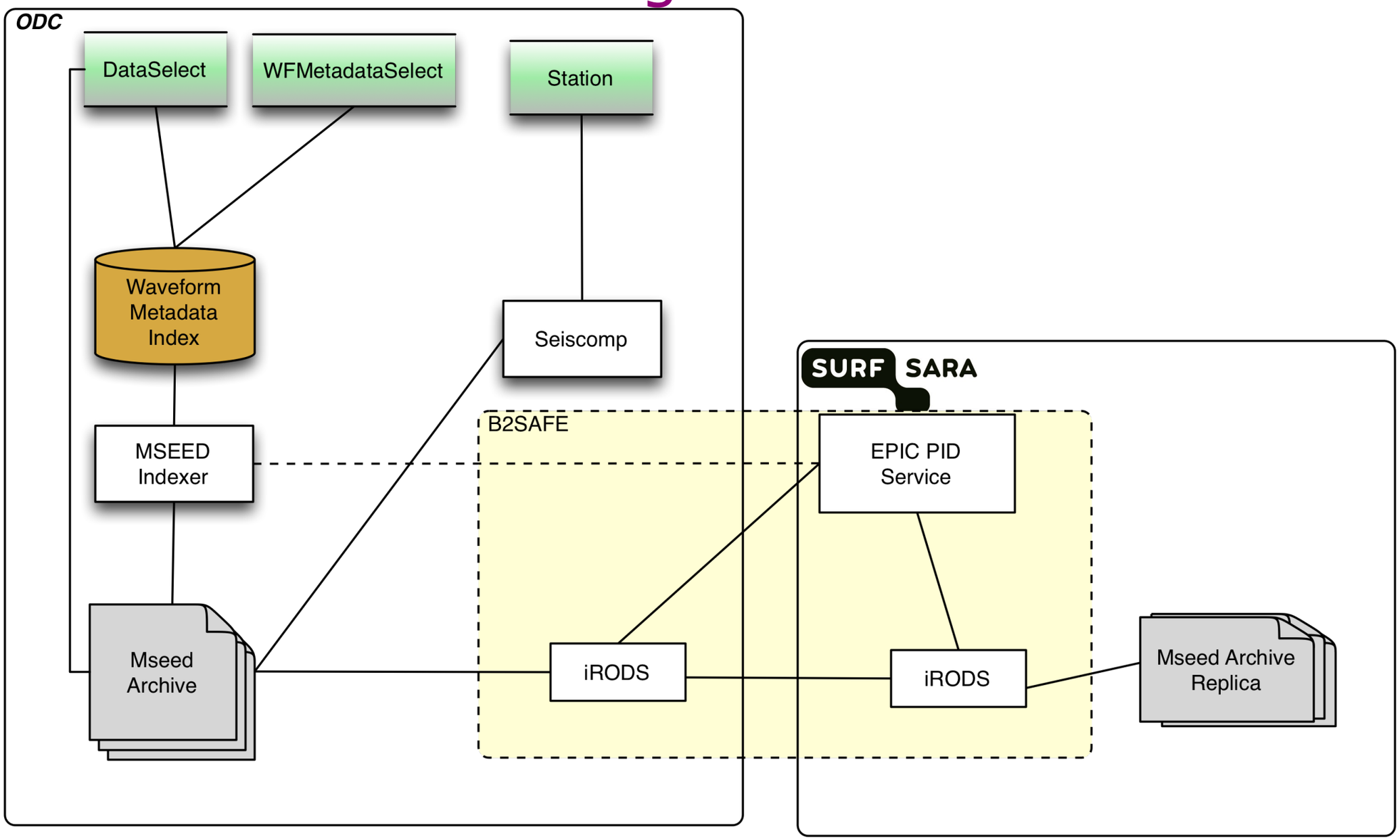
- Replication mainly for backup purposes

- The current system does not support data provenance

Impossible to trace the history of changes on a specific dataset: related to Dynamic Data

- PID not coupled with domain specific metadata

ODC Metadata Integration



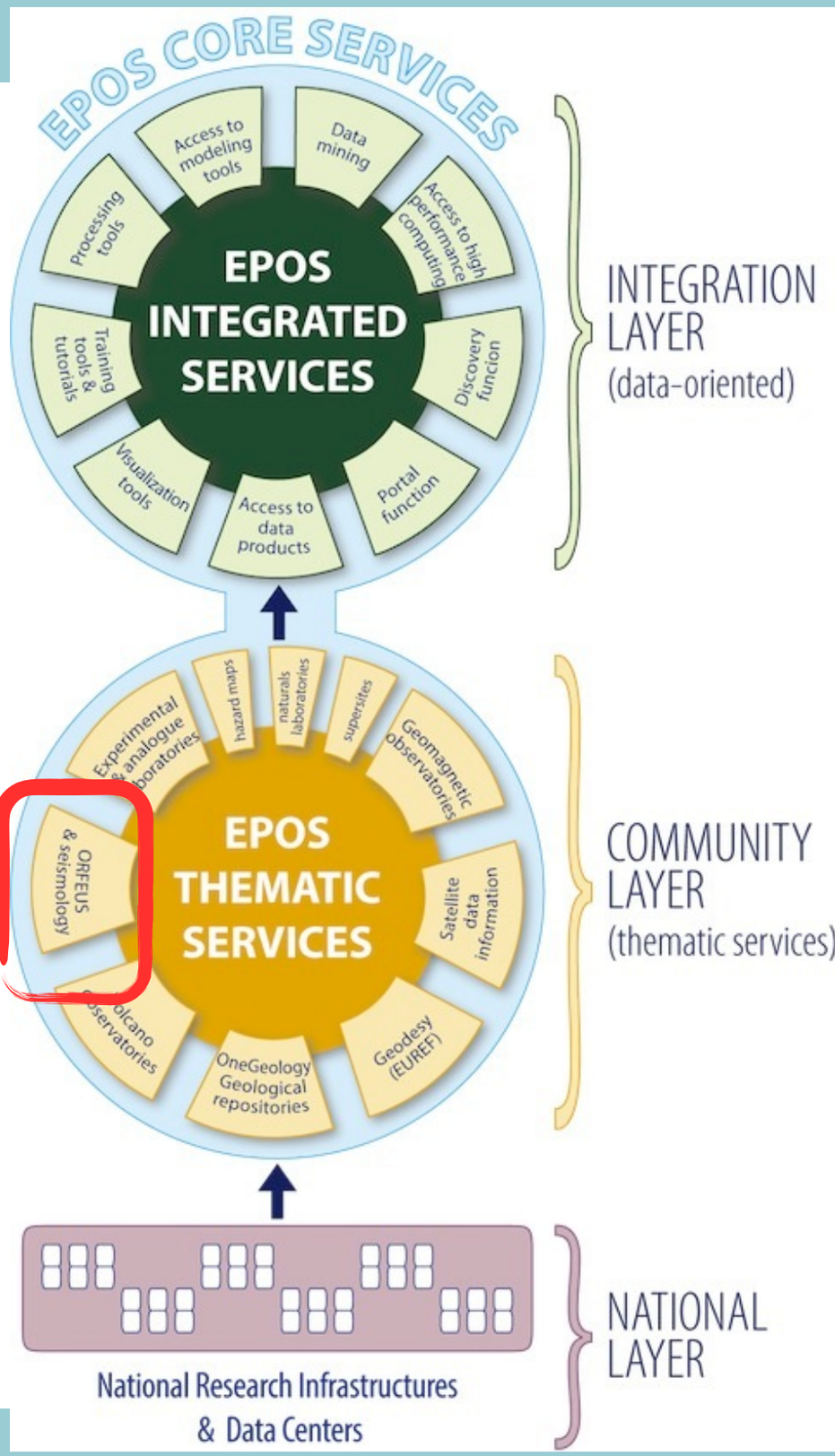
EIDA and EPOS: EPOS-S

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

European Plate Observing System

The existing national Research Infrastructures (RIs) for solid Earth science in Europe generate data and information and are responsible for the operation of instrumentation in each country. These RIs are integrated into the EPOS Thematic Services, which represent dedicated services for each specific community.

The distinct Thematic Services are further joined up to create the EPOS Integrated Services consisting of a variety of multidisciplinary services that will allow the access to data, data products, processing and visualisation tools and computational codes and resources for different stakeholders, not limited to the scientific community.



EIDA/EPOS-S plans within EUDAT2020

- Set up a robust, scalable, reliable and secure infrastructure to support the federation of datacenters

- Enrich data streams with PIDs and detailed metadata

- Enable reproducibility and provenance

- Automatic safe replication of datasets on external data resources

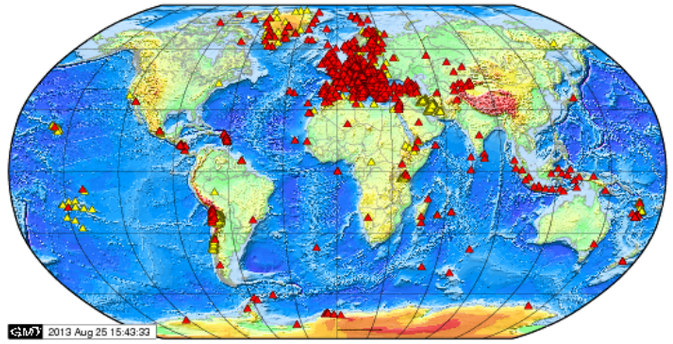
- Effective exploitation of replicas: reliability, failover, disaster recovery, computation, optimisation

- Federated Discovery and Access

- Federated Identity Management (in discussion with eduGAIN/GEANT)

EIDA in EUDAT2020

Users: Geoscientists etc...



BGR

ETH

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

GFZ Helmholtz Centre Potsdam

B2SAFE Replicate Research Data Safely

KIT Karlsruher Institut für Technologie

CINECA

B2SAFE Replicate Research Data Safely

INGV

RESIF

Orfeus

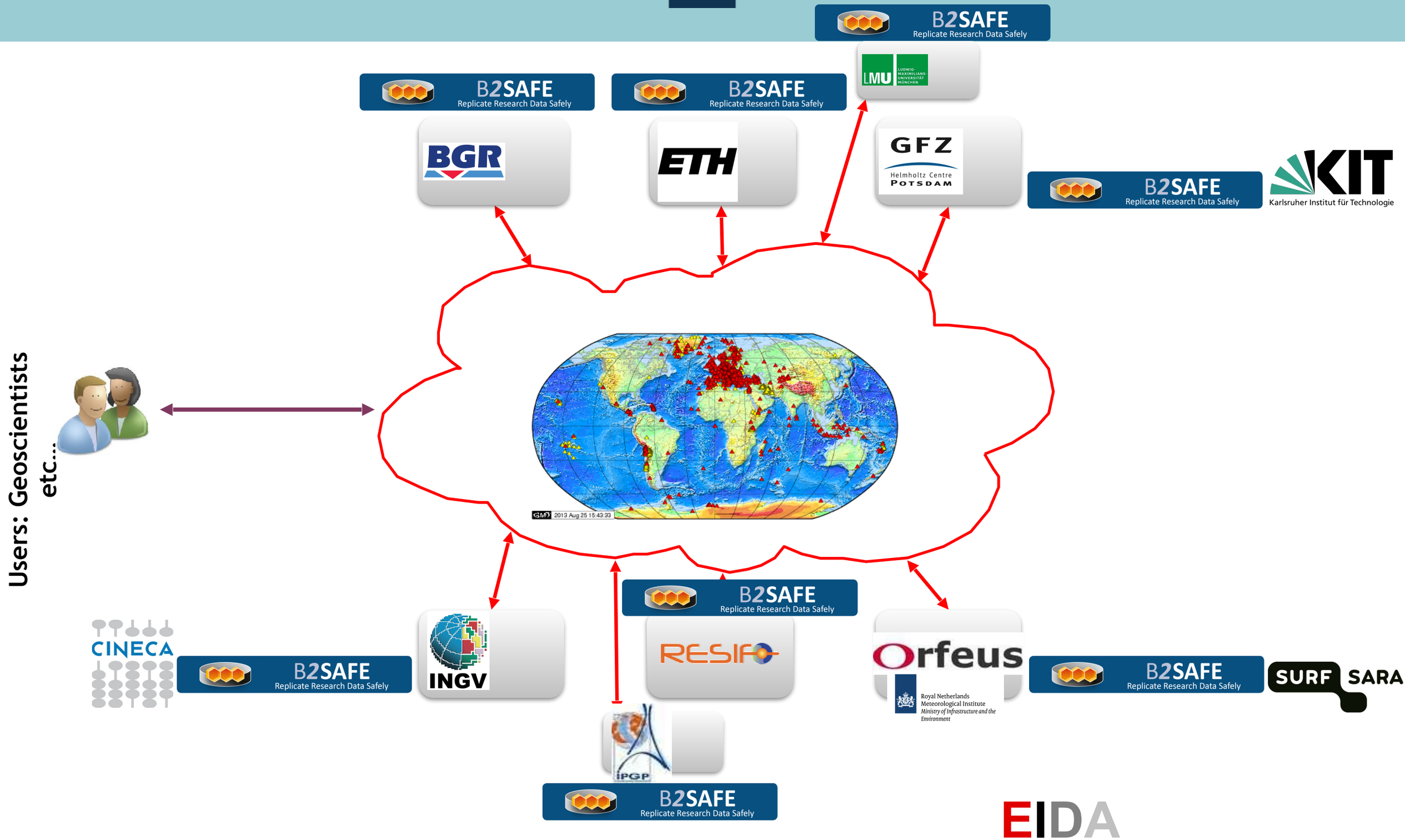
Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

B2SAFE Replicate Research Data Safely

SURF SARA

IPGP

EIDA


Users: Geoscientists
etc...
B2SAFE
Replicate Research Data Safely
BGR
ETH
LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
GFZ
Helmholtz Centre
POTSDAM
KIT
Karlsruher Institut für Technologie
2013 Aug 25 15:43:33
CINECA
INGV
RESIF
Orfeus
Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment
SURF SARA
IPGP
EIDA

B2SAFE
Replicate Research Data Safely

B2SAFE
Replicate Research Data Safely

B2SAFE
Replicate Research Data Safely

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

BGR

ETH

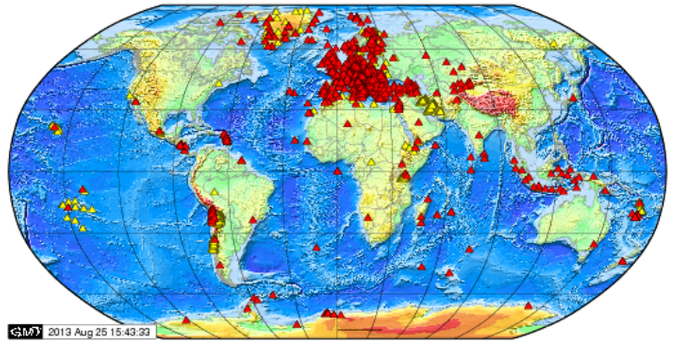
GFZ
Helmholtz Centre
POTSDAM

B2STAGE
Get Data to Computation

B2SAFE
Replicate Research Data Safely

KIT
Karlsruher Institut für Technologie

Users: Geoscientists
etc...



CINECA

B2STAGE
Get Data to Computation

B2SAFE
Replicate Research Data Safely

INGV

B2SAFE
Replicate Research Data Safely

RESIF

Orfeus

Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

B2STAGE
Get Data to Computation

B2SAFE
Replicate Research Data Safely

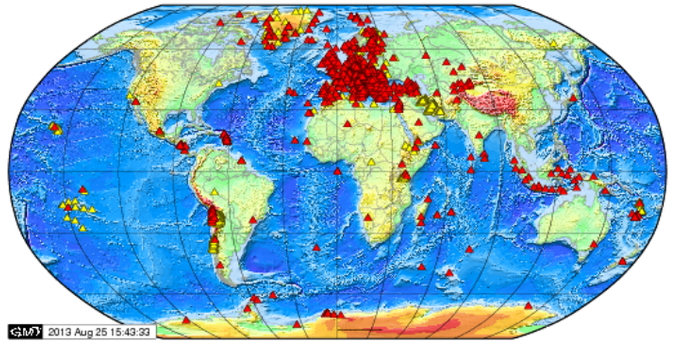
SURF SARA

IPGP

B2SAFE
Replicate Research Data Safely

EIDA

Users: Geoscientists etc...



B2STAGE Get Data to Computation
B2SAFE Replicate Research Data Safely

LMU Ludwig-Maximilians-Universität München

BGR

ETH

GFZ Helmholtz Centre Potsdam

KIT Karlsruher Institut für Technologie

CINECA

INGV

RESIF

Orfeus

Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

SURF SARA

IPGP

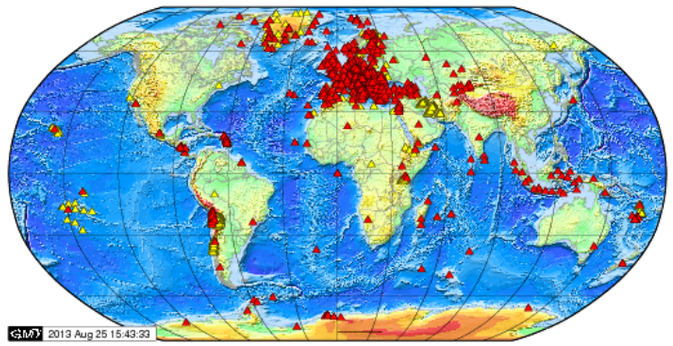
EIDA

Users: Geoscientists etc...

Data Discovery
Data Access
Data Computation
AAI
Citation
Workflows
...

B2SHARE
Store and Share Research Data

B2FIND
Find Research Data



B2STAGE
Get Data to Computation

B2SAFE
Replicate Research Data Safely

BGR

ETH

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

GFZ Helmholtz Centre Potsdam

B2?
...

KIT Karlsruher Institut für Technologie

CINECA

INGV

RESIF

Orfeus

Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

SURF SARA

IPGP

New node

EIDA

EIDA
- home
- management and structure
- contributing networks
- monitoring
- data access
- data acknowledgements

European Integrated Data Archive (EIDA)

EIDA, an initiative within ORFEUS, is a distributed data centre established to (a) securely archive seismic waveform data and related metadata, gathered by European research infrastructures, and (b) provide transparent access to the archives by the geosciences research communities.

EIDA nodes are data centres which collect and archive data from seismic networks deploying broad-band sensors, short period sensors and accelerometers. Networks contributing data to EIDA are listed in the **ORFEUS EIDA networklist**. All data from the **VEBSN** at ORFEUS Data Center are available through EIDA.

Technically, EIDA is based on an underlying architecture developed by GFZ to provide transparent access to all nodes' data. Data within the distributed archives are accessible via the **ArcLink protocol**.

EIDA data access

This map shows EIDA stations with unrestricted data access (total number: **4639**). Stations in green are open (operational; **1747**) according to the current metadata, while stations in orange are closed (**2892**). From the total stations **2021** belong to a permanent network (**80**), while **2618** stations belong to a temporary network (**59**). ORFEUS Data Center updates this map daily (last update: **2014-09-18 00:14:02**).

open station closed station EIDA node

EIDA nodes and additional services/data:

Currently 8 nodes are contributing their data to EIDA. Six of them, called "primary nodes", have committed resources to ensure the EIDA operations and to support further developments. The region indicates the focus of operation for each node. Identical requests for open data to each EIDA nodes provide identical data. Each node may also provide unique, restricted data and additional services. Click on a node for detailed information on specific datasets and/or access tools.

ODC	European-Mediterranean area (VEBSN)
GFZ	European, Global, temporary deployments
RESIF	France + Global temporary deployments
INGV	Italy, European-Mediterranean (MedNet)
ETH	Switzerland
BGR	Germany
IPGP	France (volcanological observatories) + Global (GEOSCOPE)
LMU	Germany (BayernNetz)

For more info visit us at
http://www.orfeus-eu.org/eida/eida.html

EIDA
- home
- management and structure
- contributing networks
- monitoring
- data access
- data acknowledgements

European Integrated Data Archive (EIDA)

EIDA, an initiative within ORFEUS, is a distributed data centre established to (a) securely archive seismic waveform data and related metadata, gathered by European research infrastructures, and (b) provide transparent access to the archives by the geosciences research communities.

EIDA nodes are data centres which collect and archive data from seismic networks deploying broad-band sensors, short period sensors and accelerometers. Networks contributing data to EIDA are listed in the **ORFEUS EIDA networklist**. All data from the **VEBSN** at ORFEUS Data Center are available through EIDA.

Technically, EIDA is based on an underlying architecture developed by GFZ to provide transparent access to all nodes' data. Data within the distributed archives are accessible via the **ArcLink protocol**.

EIDA data access

This map shows EIDA stations with unrestricted data access (total number: **4639**). Stations in green are open (operational; **1747**) according to the current metadata, while stations in orange are closed (**2892**). From the total stations **2021** belong to a permanent network (**80**), while **2618** stations belong to a temporary network (**59**). ORFEUS Data Center updates this map daily (last update: **2014-09-18 00:14:02**).

▲ open station ▲ closed station ■ EIDA node

EIDA nodes and additional services/data:

Currently 8 nodes are contributing their data to EIDA. Six of them, called "primary nodes", have committed resources to ensure the EIDA operations and to support further developments. The region indicates the focus of operation for each node. Identical requests for open data to each EIDA nodes provide identical data. Each node may also provide unique, restricted data and additional services. Click on a node for detailed information on specific datasets and/or access tools.

ODC	European-Mediterranean area (VEBSN)
GFZ	European, Global, temporary deployments
RESIF	France + Global temporary deployments
INGV	Italy, European-Mediterranean (MedNet)
ETH	Switzerland
BGR	Germany
IPGP	France (volcanological observatories) + Global (GEOSCOPE)
LMU	Germany (BayernNetz)

For more info visit us at
http://www.orfeus-eu.org/eida/eida.html

Thank You!