Enriching the EIDA seismic data archive with EUDAT services

EIDA: a transparent distributed data archive for seismology

EUDAT: new services for generic large scale distributed data

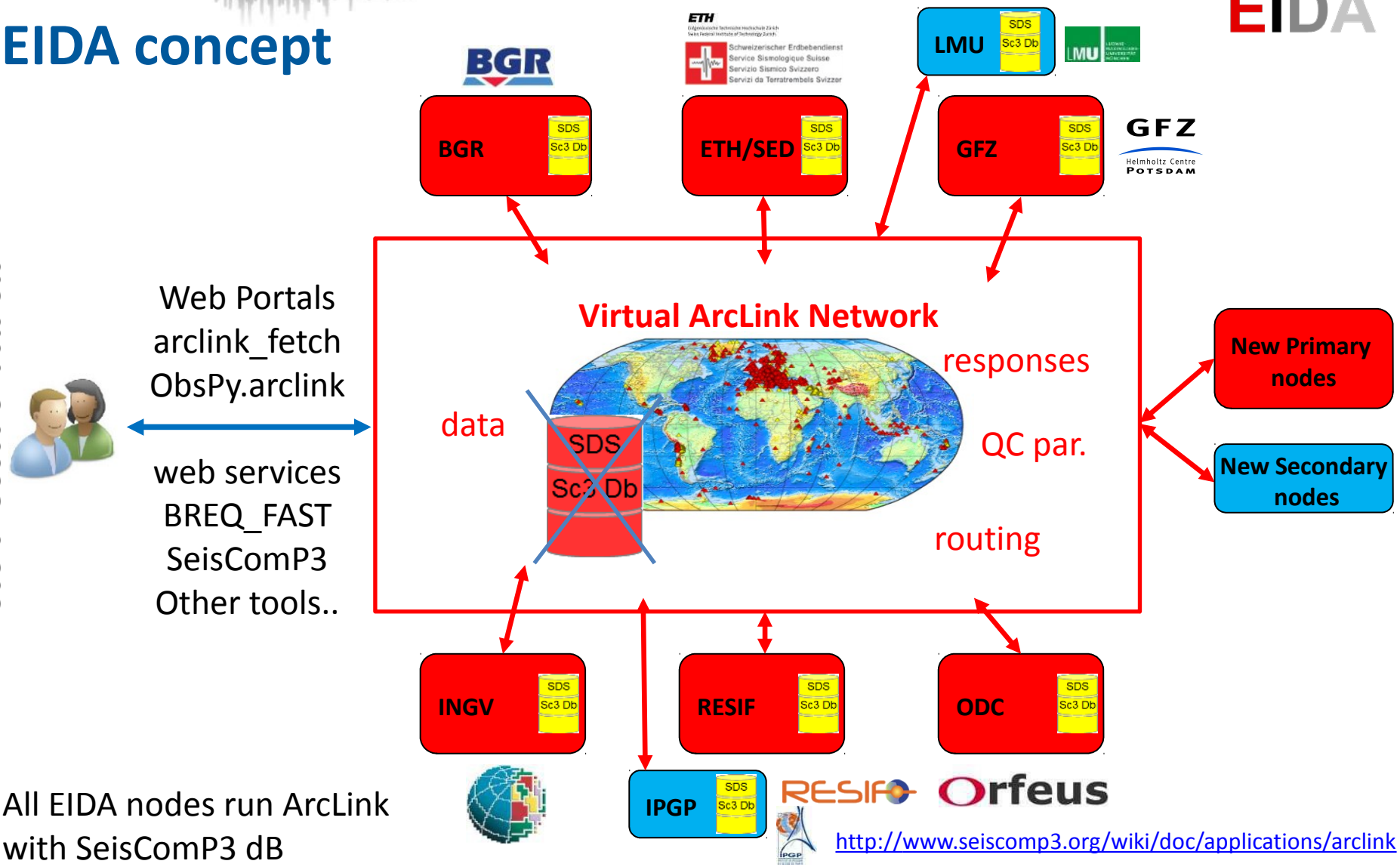
A. Strollo, L. Trani, P. L. Evans, A. Michelini,
M. Fares, J. Quinteros
GFZ / ODC / INGV

pevans@gfz-potsdam.de peter.evans@jabber.org

EIDA overview

- Basic idea: *E*uropean *I*ntegrated waveform *D*ata *A*rchive
- Federation/collaboration of data centres (“nodes”)
  - exchanging seismic metadata (“inventory”) and
  - resource location (“routing”)
- Data stays at original nodes
  - size (requests up to gigabytes)
  - access control (limited users, encrypted delivery)
  - inertia
- Rapid AND long-term availability are required.

EIDA concept



All EIDA nodes run ArcLink with SeisComP3 dB

http://www.seiscomp3.org/wiki/doc/applications/arclink

EIDA Management

- MoUs between ORFEUS (KNMI) and the nodes (Nov. 2012)
- EIDA Management Board – sets overall policy and strategy
- EIDA Technical Commission – ensures operations, review performance, coordinate modifications and extensions
- Commitment by nodes to provide open access.

Orfeus

EIDA
European Integrated Data Archive

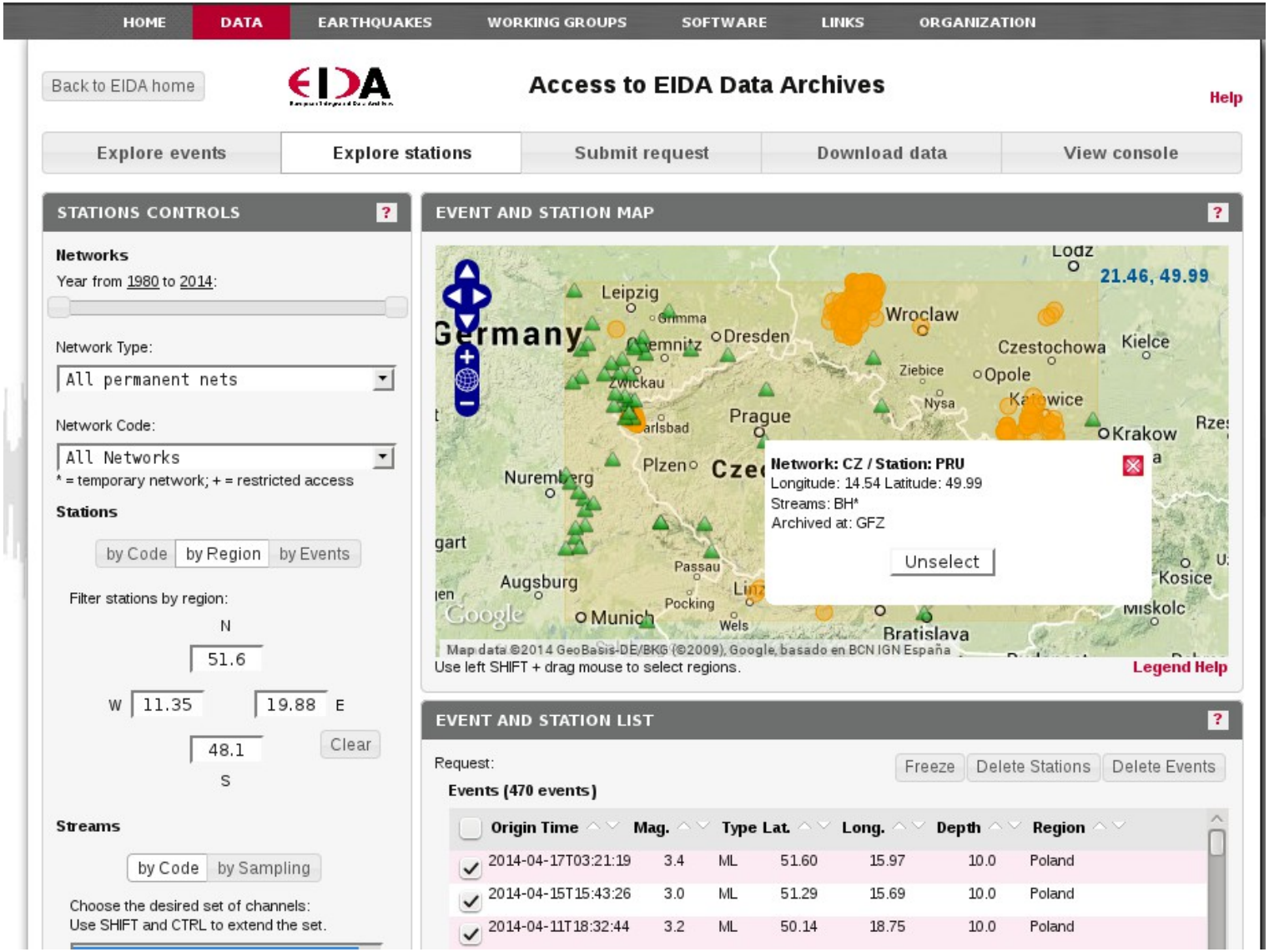
EIDA data exchange

- Waveform data is hosted at a single node; tools need routing information.
- Inventory (seismic metadata) is copied to all nodes.
- Routing table is exchanged between all nodes.
- Nightly metadata updates; one designated node is responsible for each network. 68MB as XML.

Data access

- WebDC/Orfeus portals for web-based access
  - **http://www.orfeus-eu.org**
  - **http://eida.gfz-potsdam.de**
- Other access methods
  - custom Arclink protocol, custom client arclink_fetch
  - web services (fdsnws)
  - e-mail based (BREQ_FAST)

EIDA and EUDAT



EIDA Evolution

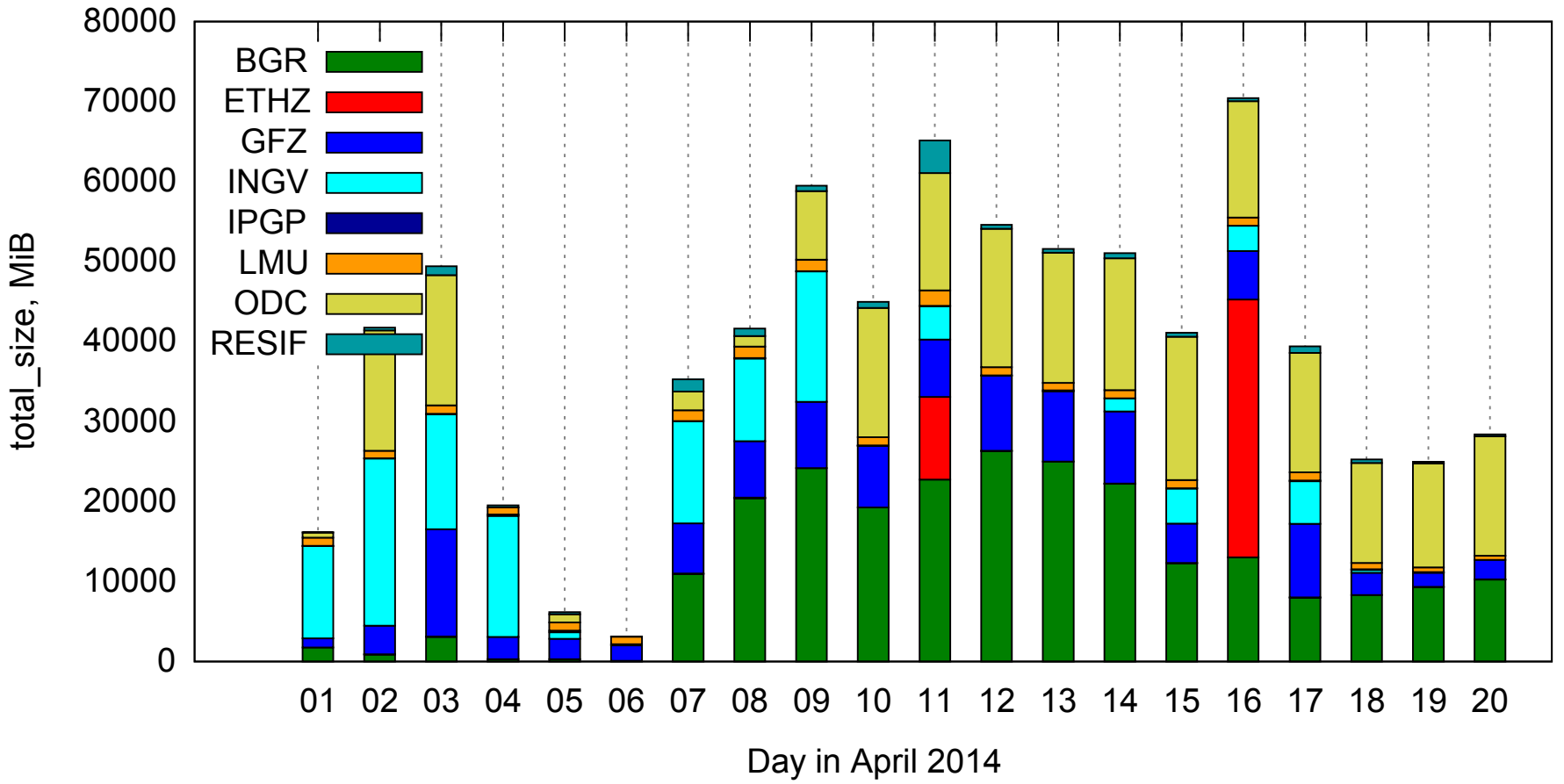
- Historically, data was held at individual small European data centres
- Each running server for custom ArcLink protocol (GFZ)
- WebDC.eu portal running since 2004; major upgrade Oct 2013.
- Beginning to offer data, metadata as web services.
- Now 8 nodes. Added 2 data centres in last 2 years.
- Today: 75 permanent, 44 temporary networks. 3958 stations. 33593 streams 0.1-100 sps: 10s of GB new data per day
- Downloads 25-50GB/day (peak days ~60GB); >200000 requests per day; 10s-100s of IP addresses per day.

EIDA and EUDAT

This map shows EIDA stations with unrestricted data access (total number: 3503). Stations in green are open (operational; 1542) according to the current metadata, while stations in orange are closed (1961). From the total stations 1978 belong to a permanent network (75), while 1525 stations belong to a temporary network (42). ORFEUS Data Center updates this map daily (last update: 2014/04/21 09:11).

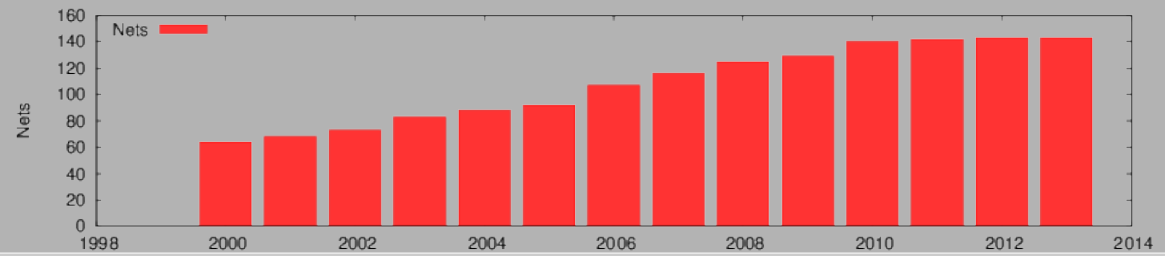
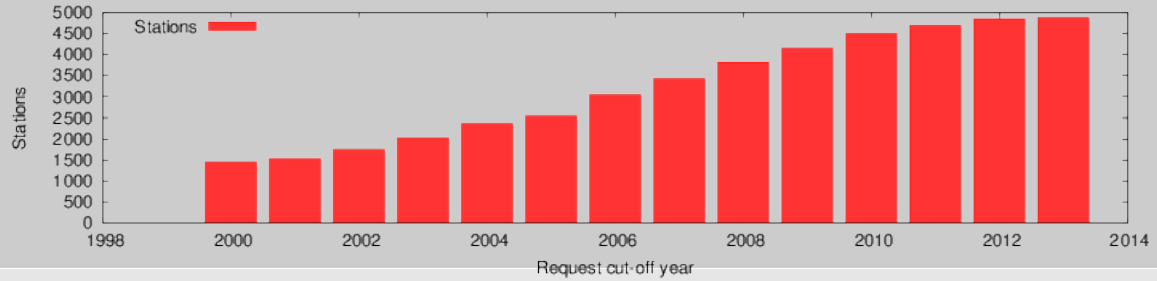
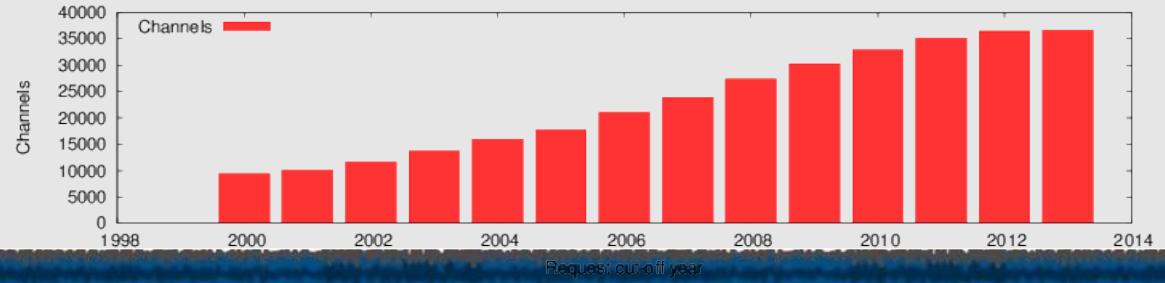
open station closed station EIDA node

Requested data volume April 2014



EIDA development

- Continuing growth in terms of nodes, networks, stations, channels, demand for parameters

Networks +7% p.a	
Stations +10% p.a.	
Channels +11% p.a.	

EIDA Future needs

- Assured secure storage
- Performance, especially given web services
- Better information about:

  - availability

  - waveform quality (noise, gaps, timing, …)

  - other parameters of interest
- Selection based on events, parameters (*discoverability*)

Enriching EIDA with EUDAT

- Started with safe replication with B2SAFE
- With local (=national?) data storage partner

  - GFZ, BGR, LMU – KIT (de)

  - IPGP, RESIF – CINES (fr)

  - INGV – CINECA (it)

  - ODC – SARA (nl)

  - SED (ETHZ) - **???** (ch)
- What if there's no national partner?
- New nodes in Turkey and Romania are expected in 2014

Enriching EIDA with EUDAT (2)

- Rich persistent identifiers

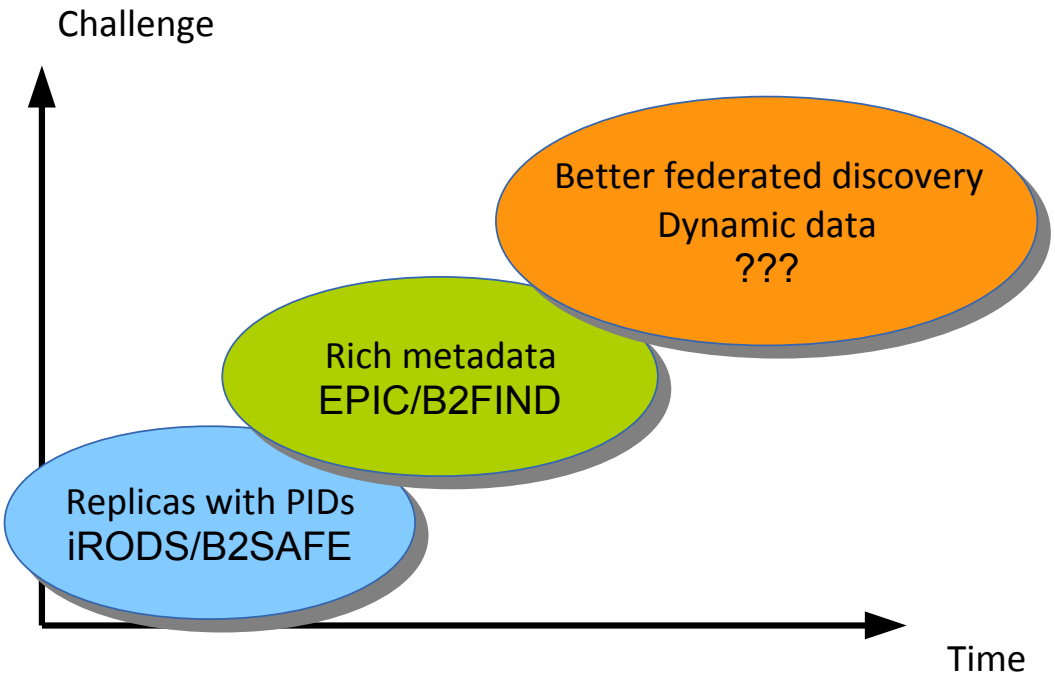
  -manage all those copies?

- Support fine-grained citation of e.g. all data used in analysis of one earthquake, or many. One per stream per day = 10^7 PID/year.

  - Will iRODS, B2SAFE etc., B2* scale?

- How to aid discovery: richer metadata?
- Use replicas to lower latency?

EIDA-EUDAT mud map



Conclusion

- EIDA already manages significant data, and the systems work “well enough” today, but...
- Room for improvement

  - reliability (B2SAFE, off-centre backup)

  - managing local copies
- Expansion needs suggest a rethink
- Use EUDAT standard services where they exist and offer value
- Especially: tools to explore the data in new ways - B2FIND
- Compute quality measures with workflow tools??