Orfeus

Promoting Observational Seismology
in the Euro-Mediterranean Region

Observatories and Research Facilities for European Seismology

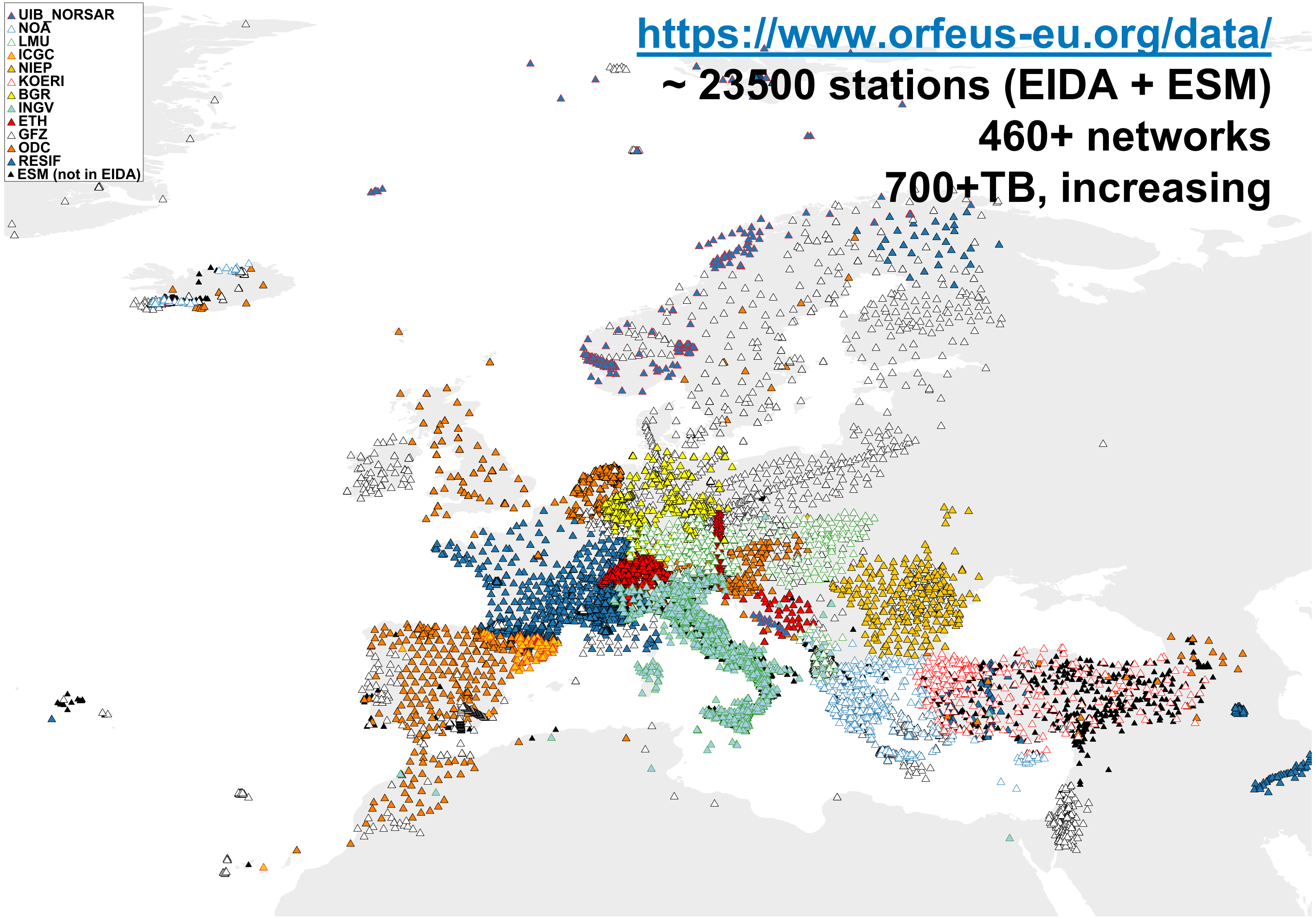
Next Generation
European Coordinated Earthquake "Datalakes"

Carlo Cauzzi (ORFEUS & SED@ETHZ),
Lucia Luzi (INGV), Susana Custodio (FCUL),
Dino Bindi (GFZ), Jarek Bienkowski (KNMI/ODC),
Eser Çakti (KOERI), Alberto Michelini (INGV),
Francesca Pacor (INGV), Javier Quinteros (GFZ), Jonathan Schaeffer (RESIF),

& other members of the ORFEUS community.

Premise

- ORFEUS is not one datacenter. ORFEUS is a federation of data centers and a community of observatories & scientists.
- ORFEUS does not have yet a formal program on computational seismology, but many ORFEUS-associated data centers are involved in relevant projects and efforts at national and European level (e.g., DT-GEO).
- ORFEUS input to this WS deals with data management strategies and standardisation.



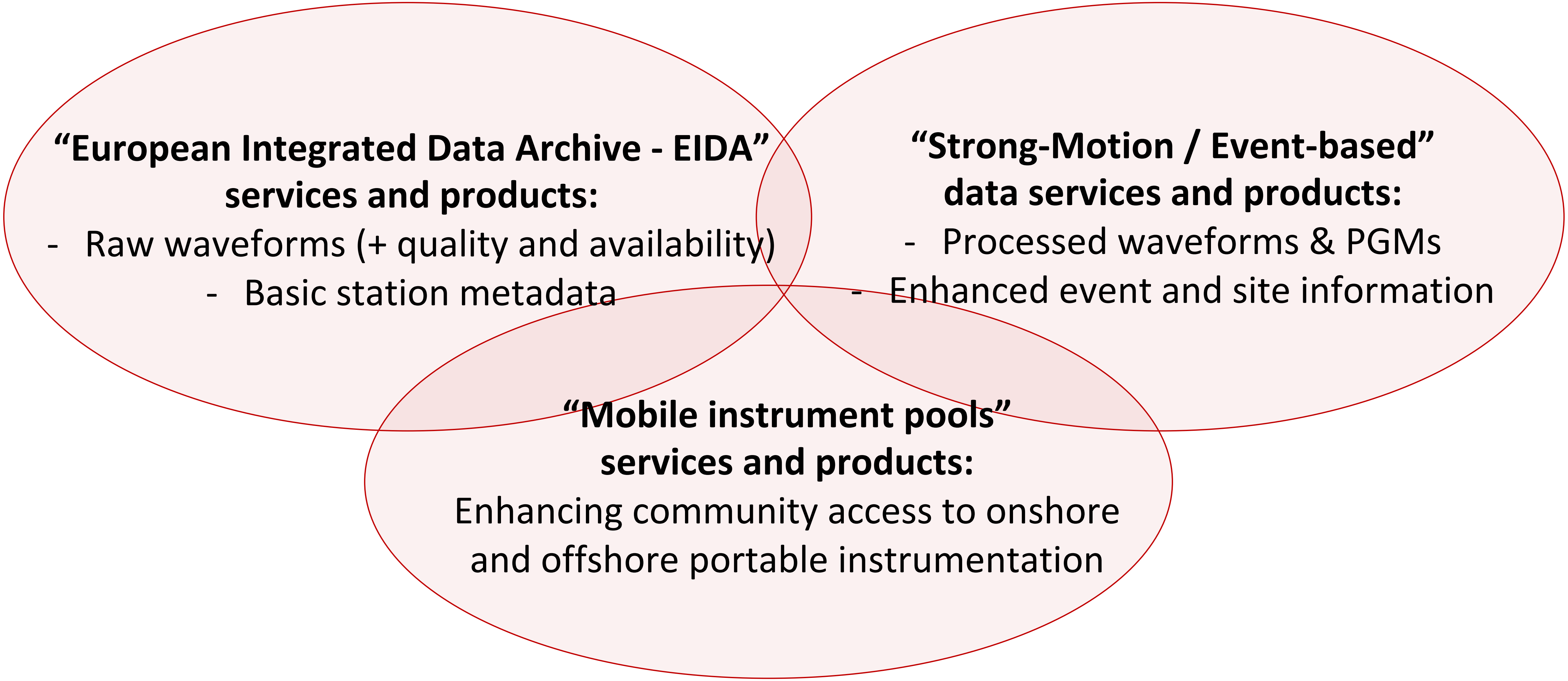
https://www.orfeus-eu.org/data/

~ 23500 stations (EIDA + ESM)

460+ networks

700+TB, increasing

Current ORFEUS service domains



“European Integrated Data Archive - EIDA” services and products:

- Raw waveforms (+ quality and availability)
- Basic station metadata

“Strong-Motion / Event-based” data services and products:

- Processed waveforms & PGMs
- Enhanced event and site information

“Mobile instrument pools” services and products:

Enhancing community access to onshore and offshore portable instrumentation

The European Integrated Waveform data Archive EIDA

Webinterface

Graphical Interface for waveform and metadata access.

Webservices

APIs for data and metadata access.

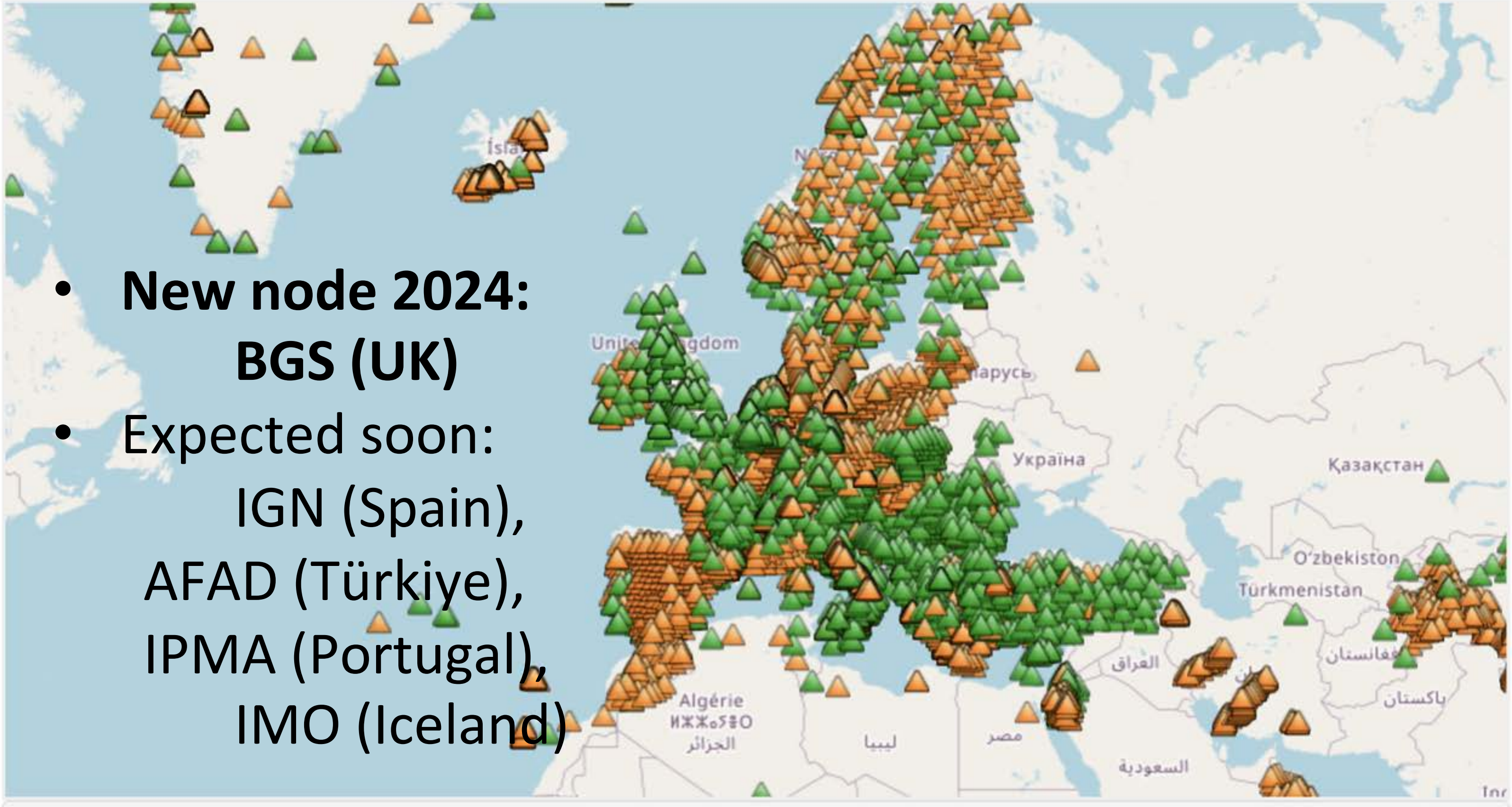
Data Quality

Interfaces for data quality visualization.

Station Book

Access to the entire EIDA station inventory.

Global collaborations: FDSN, EarthScope, etc .



- **New node 2024: BGS (UK)**
- Expected soon: IGN (Spain), AFAD (Türkiye), IPMA (Portugal), IMO (Iceland)

- ~ 23,050 stations (~4,700 in operation)
- BB, SM, SP, OBS, infrasound, gravity, DAS …
- Focus on Euro-Med, yet global experiments included
- ~135 permanent networks; ~330 temporary networks -> supplier letters to EIDA nodes -> MoU with ORFEUS
- https://www.fdsn.org/networks/citation/

EIDA/ORFEUS data access:"today"

The "pillars" of data access:
Web Interfaces for data discovery (and seldom access): e.g., https://orfeus-eu.org/webdc3/
Web Services for programmatic, frequent data access: https://www.orfeus-eu.org/data/eida/webservices/

- **FDSN webservice fdsnws-dataselect**
  Provides waveform data in MSEED format
  https://www.orfeus-eu.org/fdsnws/dataselect/1/query

- **FDSN webservice fdsnws-station**
  Provides station metadata in XML and text format.
  https://www.orfeus-eu.org/fdsnws/station/1/query

- **FDSN webservice fdsnws-availability**
  Provides data availability for channels and time windows.
  https://www.orfeus-eu.org/fdsnws/availability/1/query

- **EIDA webservice eidaws-routing**
  Provides routes to different EIDA data and services
  https://www.orfeus-eu.org/eidaws/routing/1/query

- **EIDA webservice eidaws-wfcatalog**
  Provides metadata and quality parameters of wf data
  https://www.orfeus-eu.org/eidaws/wfcatalog/1/query

- **EIDA Federator**
  Provides a single, unified access point to the entire EIDA
  http://federator.orfeus-eu.org/fdsnws/station/1/query
  http://federator.orfeus-eu.org/fdsnws/dataselect/1/query
  http://federator.orfeus-eu.org/eidaws/wfcatalog/1/query

- **ObsPy**

- **Jupyter notebooks**

- **fdsnws_scripts**

- **Usage examples available at** https://www.orfeuseu.org/data/eida/webservices/examples/workflow

EIDA/ORFEUS data access & management: “next generation”

In coordination with EarthScope + other regional institutions & FDSN, driven by Big Data and ML needs

- Web interfaces for data discovery
- New standards & web services for programmatic and massive data access:

  - new station metadata standard (json format) to allow proper description of dense/exotic experiments
  → from DAS-RCN outcome to FDSN standard

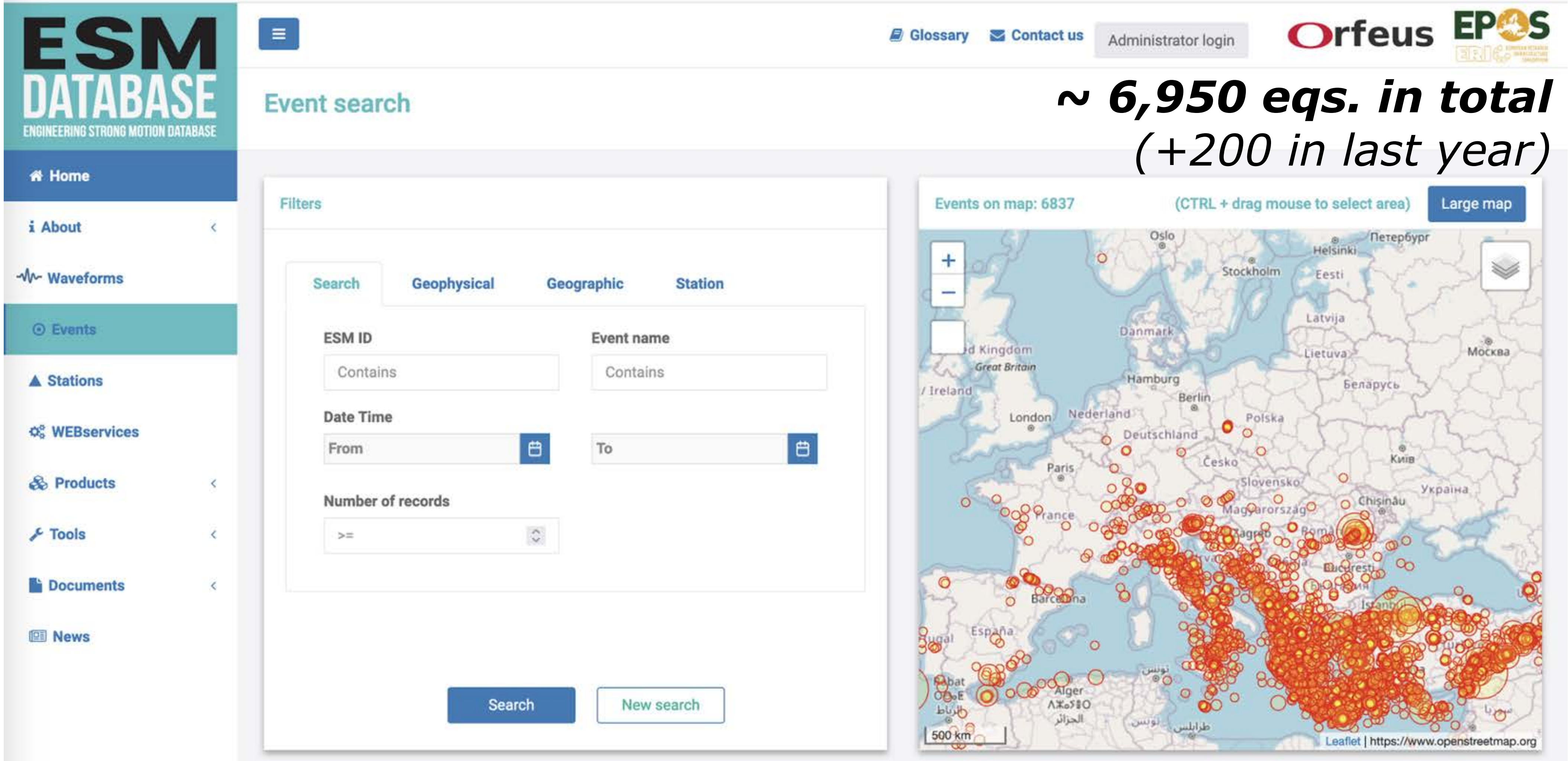
  - TileDB as new data management system for seismological data, will facilitate advanced applications by allowing integration of data, code, and processing in a single product
  → demo at ORFEUS workshop for Large-N, possibly FDSN standard

  - Asynchronous mode service for data distribution, in essence a new “dataselect” flavour pointing to S3 buckets for massive download or local processing
  → FDSN standard

  - common Authentication, Authorization Identification (AAI) approach
  → FDSN standard

- Harmonised data policies (licences), citation & FAIRification strategies

The Engineering Strong-Motion Database ESM: “today”



Event search

Filters

ESM ID: Contains

Event name: Contains

Date Time: From / To

Number of records: >=

Events on map: 6837

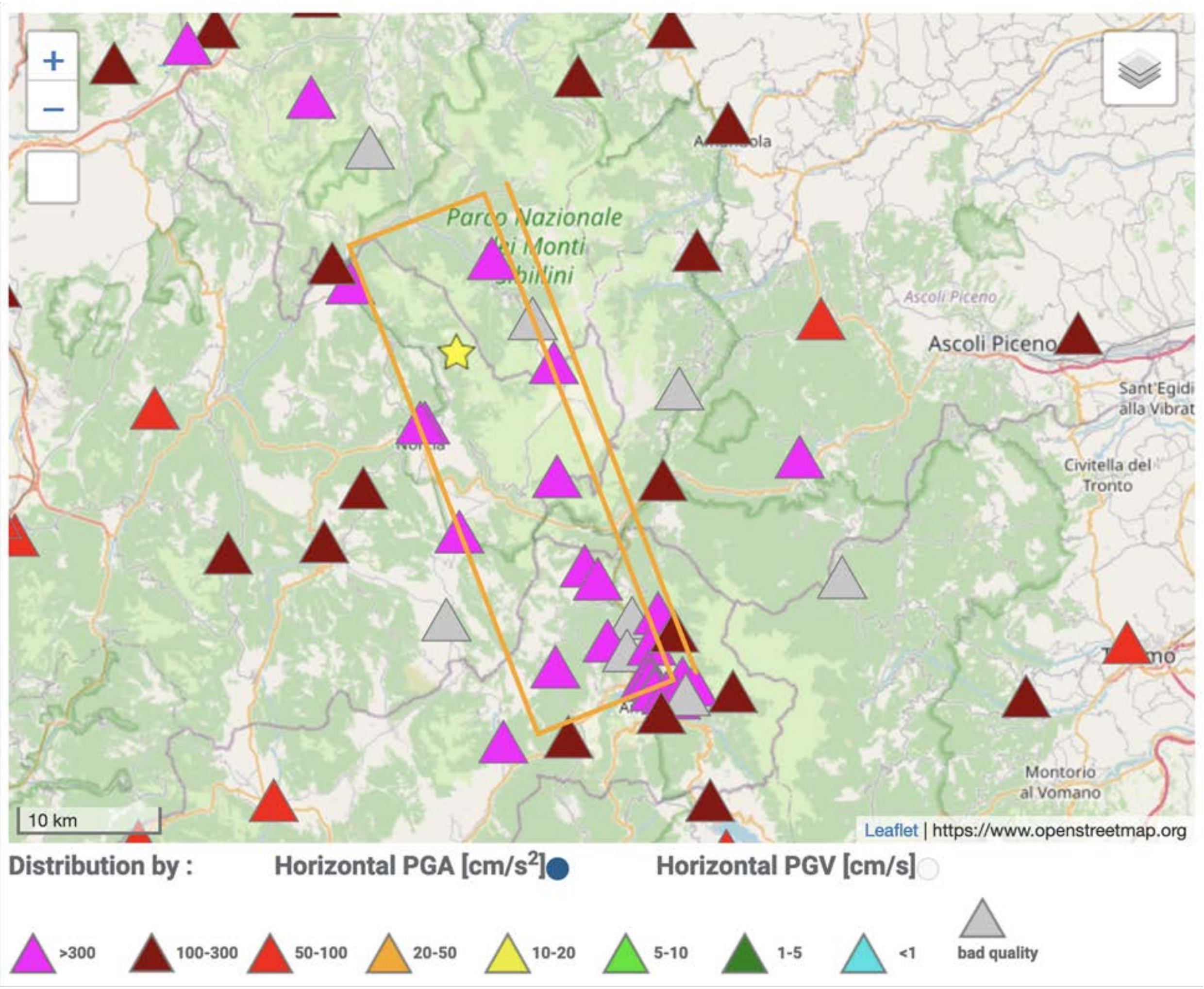
(CTRL + drag mouse to select area)

~ 6,950 eqs. in total

(+200 in last year)

Luzi et al. (SRL 2016); Lanzano et al. (SRL 2021)
Manual processing, yet transitioning to automated, Mascandola et al. (SRL 2023)

The Engineering Strong-Motion Database ESM: “today”



- First EMSC magnitude and location, updated using the best manual solution, including finite fault if available; **M (any scale) >= 4**

- Input data from EIDA & other Euro-Med SM datasets = **~4150 SM stations (1950 not yet in EIDA)**

- Delivers PGA, PGV, selected spectral ordinates, response spectra, engineering parameters, raw and and processed waveforms - also spectrum-compatible (REXELWeb) - in engineering formats including ASCII and SAC and data container like ADSF

- **Web Services** deliver manually revised input files to USGS-style ShakeMap (event, peak-motion and fault data), raw and processed waveforms, event and peak-motion information; spectrum-compatible wf - in various formats https://esm-db.eu/#/data_and_services/web_services

The Engineering Strong-Motion Database ESM: "next generation"

The "pillars" of data access:

- Web Interfaces for data discovery
- Web Services for programmatic (and massive) data access

Already discussed and agreed within ORFEUS Strong-Motion Committee:

- integration of on-scale recordings of velocity stations
- lower magnitudes
- shift towards trustworthy automated waveform data processing (sustainability, objectivity, homogeneity, quality)

To be discussed within ORFEUS Strong-Motion Committee:

- How to link and/or expose digital twins? Among the options are:
    - for relevant events, provide access to standardised repositories (e.g., S3 buckets) based on literature articles;
        (- for relevant events, allow access to selected synthetics at actual station locations, thus allowing generation of flatfiles of synthetic and recorded data)
        (- synthetic flatfiles only)

In all cases, besides standardisation of formats, community agreements on model authoritativeness / reliability is needed.

Towards concluding

Review of thoughts / suggestions provided by ORFEUS in 2022

**(which informed to a large extent
the executive summary of the first COSMOS
workshop on simulations),**

where do we stand today?

Some thoughts / suggestions: 1 – conduct a landscape analysis

Map the community “de-facto” standards and needs with a survey:

- understand if the community is ready for open and FAIR* data sharing – if not, embark the long journey to convince them: no need to plan archival and dissemination if there is no culture of open data sharing, [and no international funding (e.g., EC) without FAIR principles]
- reach out to both users and providers – are these different stakeholders?
- map data & metadata formats being used
- map data & metadata types (including flat-files and ML datasets) being used
- map software being used for pre- & post-processing, and for the simulations
- **map the needs of users** in terms of access and post-processing

* FAIR = findable, accessible, interoperable and reusable

DONE via COSMOS, need to repeat?

Some thoughts / suggestions: 2 – data or products?

Based on “the needs of users in terms of access and post-processing”:

- place emphasis on expected data usage

- need to deliver data (synthetic waveforms & associated metadata) or data products (e.g., IMs on x,y,z grids, data products that allow access to selected subsets)? This is crucial to start a discussion on (meta)data types and format standardization

DONE via COSMOS, need to repeat?

Some thoughts / suggestions: 3 – don’t move the data

Keep the data where they are already archived:

- Usually, the active research institutions in numerical modelling have good plans for data storage and back-up; transferring the data to other institutions to distribute them is in principle not necessary and might simply multiply the storage costs (I acknowledge that the US perspective might be different due to the role of the IRIS DMC, that is already distributing selected synthetics)

- **Promote strategies to access the data at the ”owner” data center**

- Learn from the HPC world: consider processing data where they are, download only post-processing results

Still very valid

Some thoughts / suggestions: 4 – consider existing standards & formats

- Consider popular file formats and “containers” (still standardization of the content needed)

- If possible, rely on existing standards for data formats and disseminations: for example, if the synthetic waveforms could be downsampled in space and time and provided in miniSEED, and the metadata in StationXML, you could use the existing FDSN standard webservices for data dissemination … or existing SM-community approaches

- **Look for advice from Seismology, Computational Seismology, ML & HPC communities (e.g., SCEC, ChEESE, etc…)**

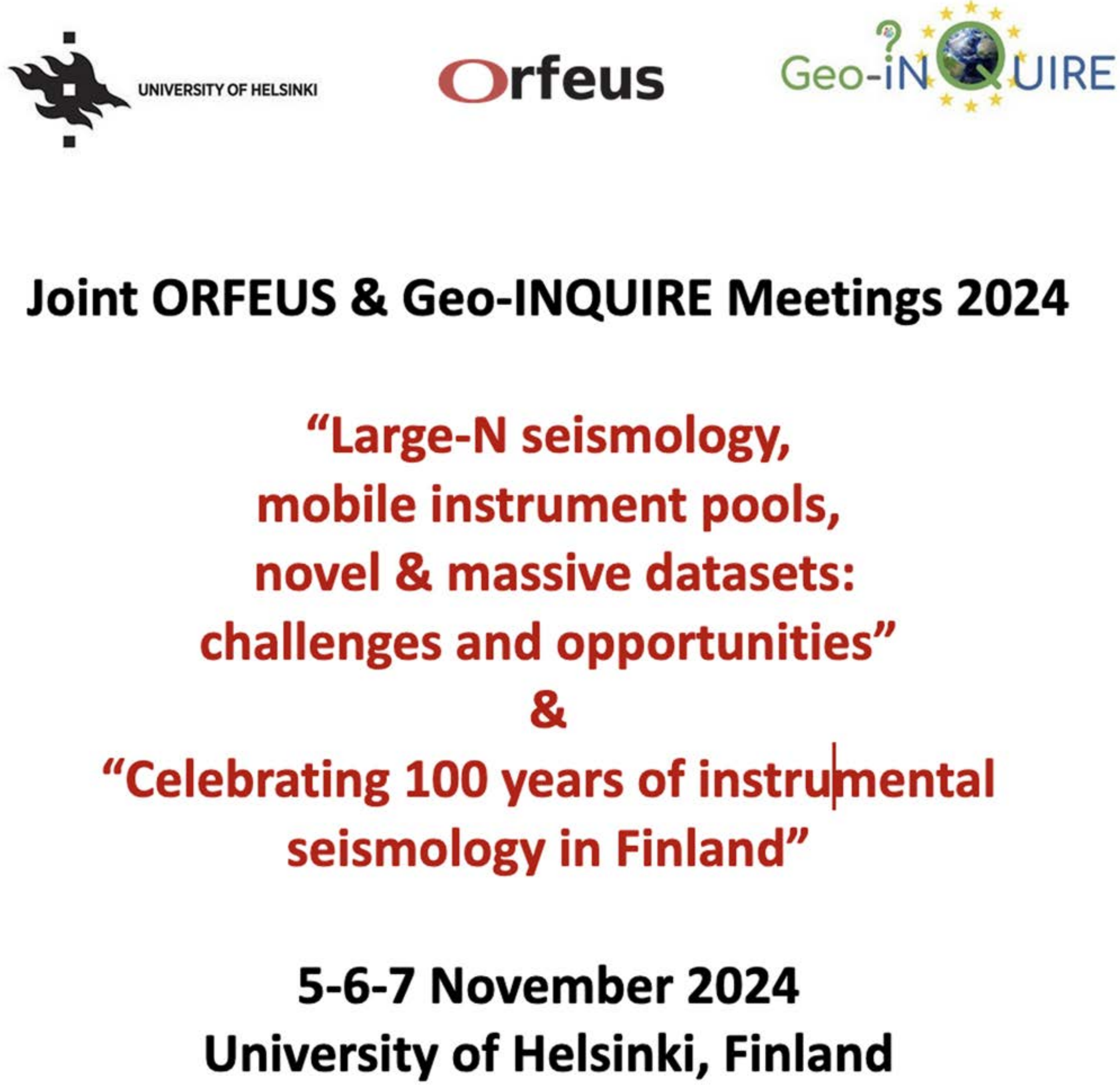
Still very valid + monitor how the data management landscape is evolving

Some thoughts / suggestions: 5 – towards standardisation and guidelines

- make the goal & mandate clear: why is this needed? who benefits? why now? who should lead and why?
- advertise the initiative
- **think global & federated**: “simulation groups” are spread all around the world, ensure broad outreach;
- **think inclusive**: don’t limit your contacts to the engineering community; most of the data management know-how on this topic is actually in the **CompSeis, ML and HPC communities**
- **think modern**: web interfaces are meant for data discovery and occasional access; routine access to data is via webservices and APIs (not to mention cloud-based approaches); consider data formats designed for large volumes and exotic datasets
- **look for consensus**: leveraging on existing de-facto standards, develop guidelines that can be adopted / implemented with minimal o/h.

Stronger community engagement/involvement/acknowledgment is needed

BTW, speaking of data management … ORFEUS Annual Workshop 2024



UNIVERSITY OF HELSINKI Orfeus Geo-iNQUIRE

Joint ORFEUS & Geo-INQUIRE Meetings 2024

**"Large-N seismology,
mobile instrument pools,
novel & massive datasets:
challenges and opportunities"
&
"Celebrating 100 years of instrumental
seismology in Finland"**

**5-6-7 November 2024
University of Helsinki, Finland**

By Tapio Haaja (unsplash.com)

A few seats are still available, especially for ECS, females and colleagues from Horizon Europe widening countries.