

EMSO Data Management Strategy

Data Management Service Group

Enoc Martínez (UPC)

EMSO Strategic Workshop

Rome, 11-13th March 2025



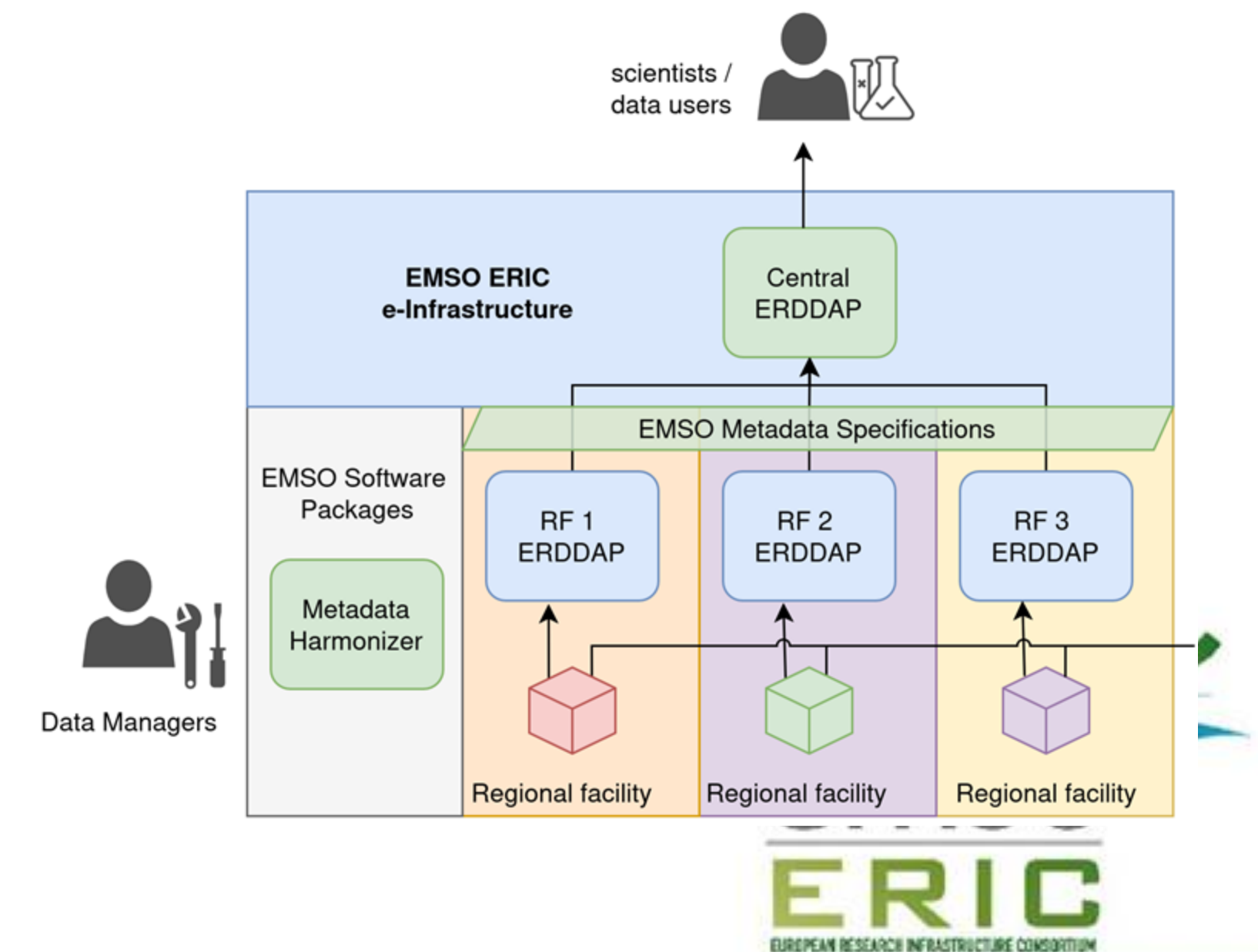
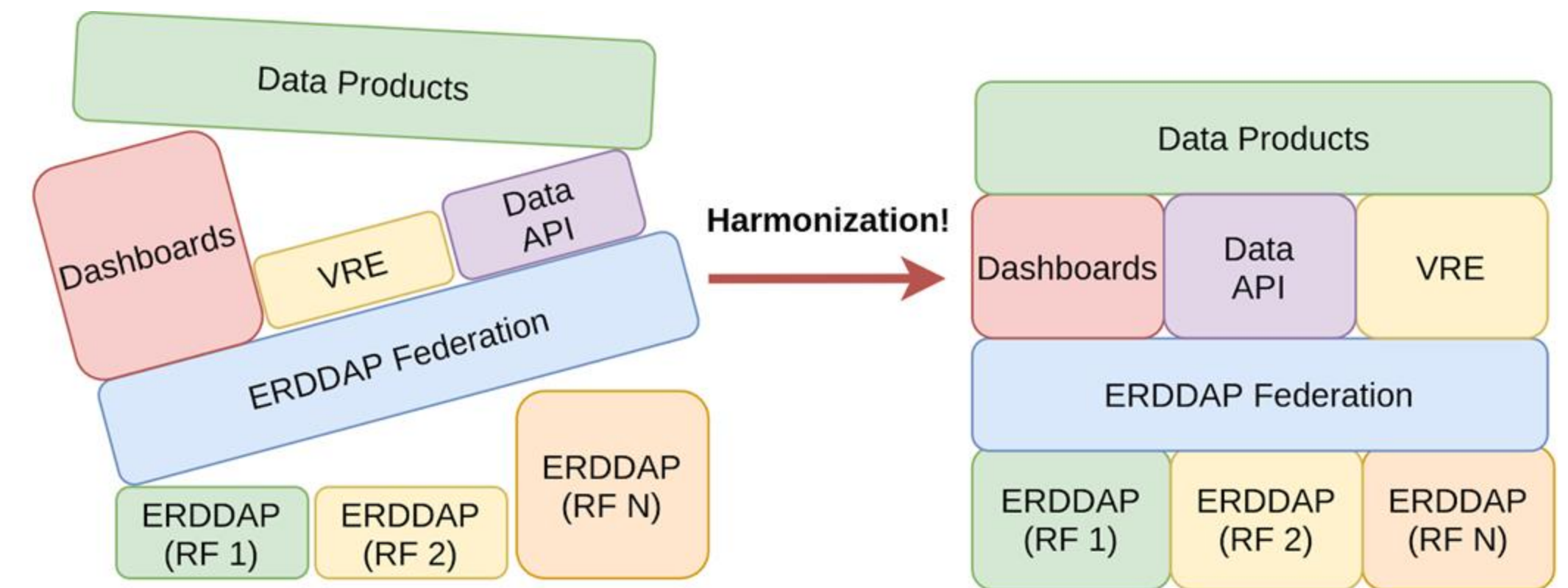
Current EMSO IT Status

EMSO Metadata Specifications

- Rules to encode data & metadata
- Built on existing initiatives: Climate & Forecast, NVS, OceanSITES, SeaDataNet...

ERDDAP Federation

- Single point of access to EMSO data
- Human and machine-actionable interfaces



Current EMSO IT Status

Harmonizer Toolbox

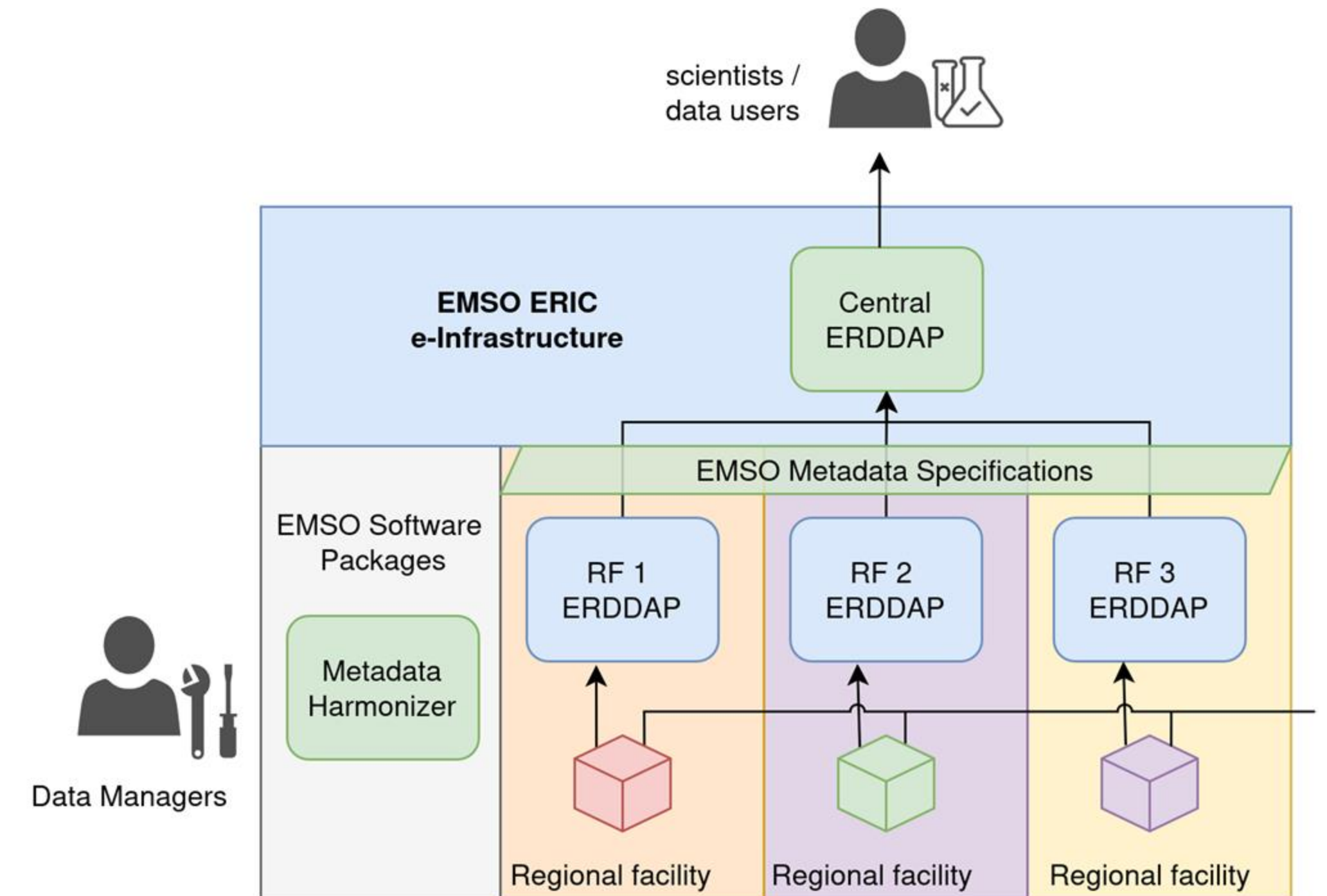
- **Dataset Generator:**
 - Creates EMSO-compliant NetCDF datasets
- **Metadata Checker**
 - Checks the metadata in datasets, providing a compliance score
- **ERDDAP autoconf**
 - Automatically dataset integration to ERDDAP
 - Avoids painful XML configuration



EMSO IT Services

Planned IT Infrastructure

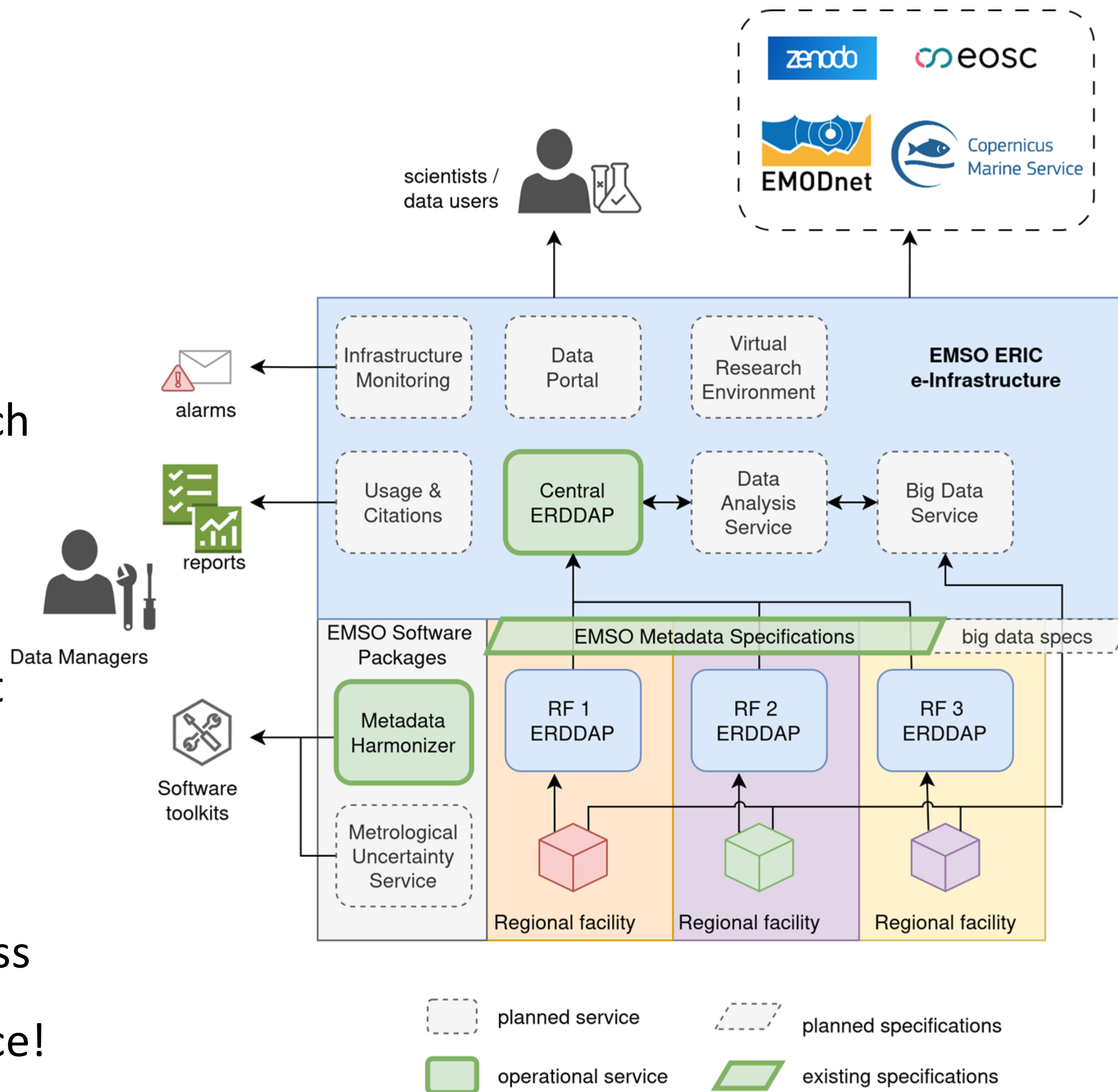
- Current IT infrastructure not enough!
- Better data access to users / aggregators
- Monitor data impact / usage
- Provide tools to internal data managers
- Handle more data types (not only time series)
- Foster new science



EMSO IT Services

Planned IT Infrastructure

- Maintain federated data approach
- Focus on internal/external users
- Internal:
 - Tools to facilitate management
 - Reports / alarms
- External users:
 - Better, easier, faster data access
 - Data Analysis Services -> science!



EMSO IT Services

EMSO Vocabulary

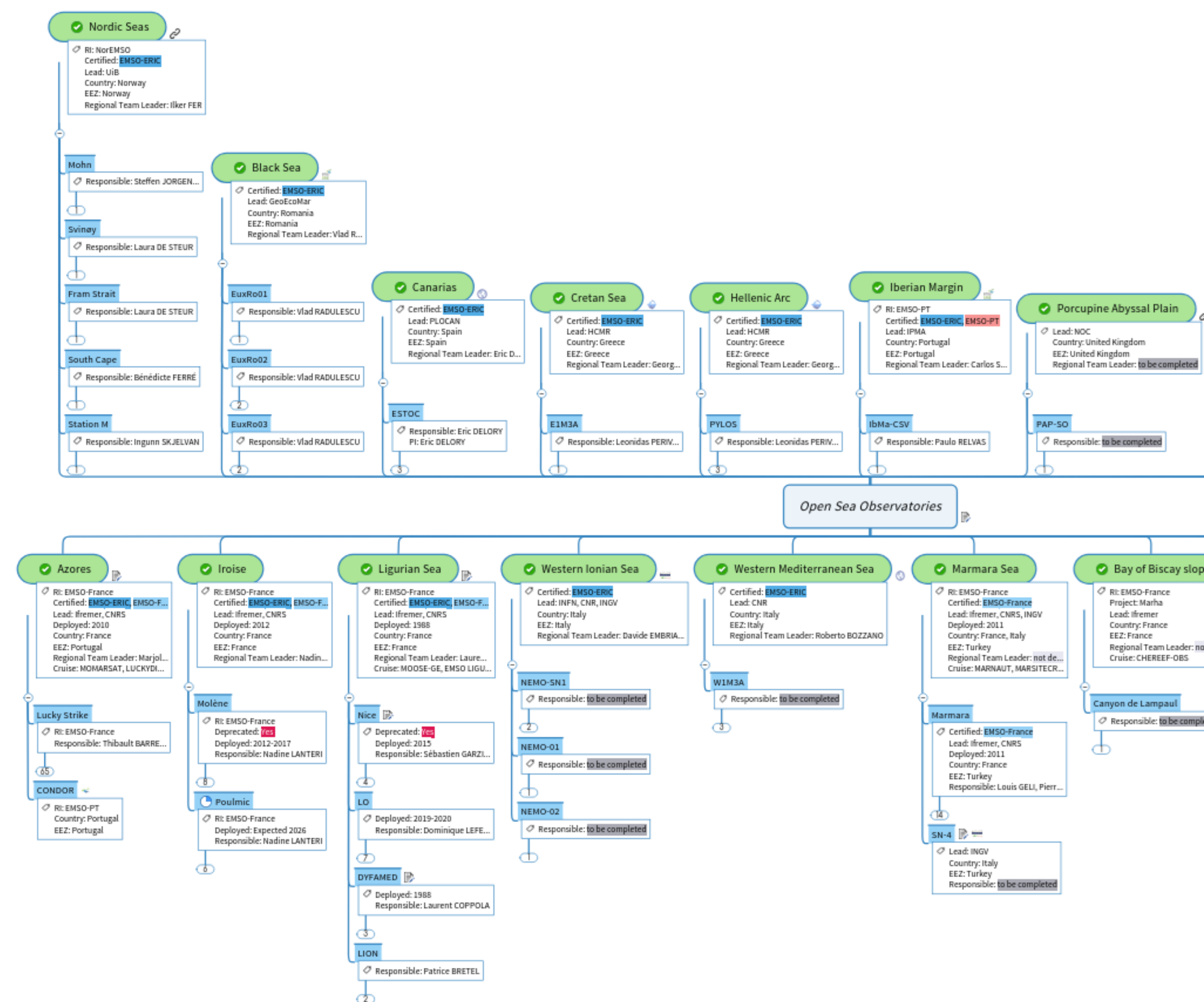
- Machine-actionable hierarchical vocabulary:
 - Regional Facility
 - Site
 - Platform
- Building block for future services

Priority: **HIGH**

Development Cost: 1 PM

Operational Cost: 0.5 PM

Target users: internal / external



EMSO IT Services

GoAccess Reporting

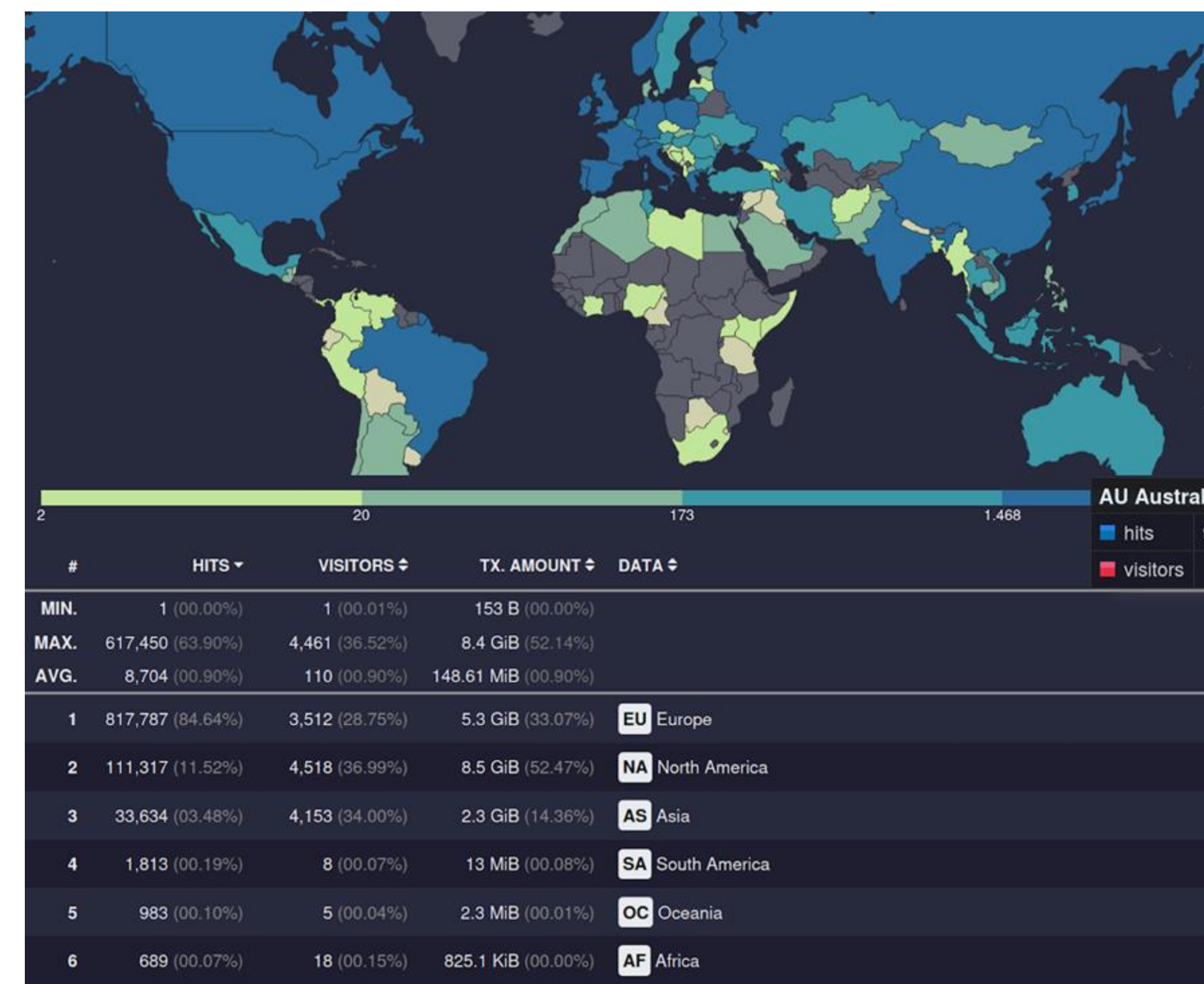
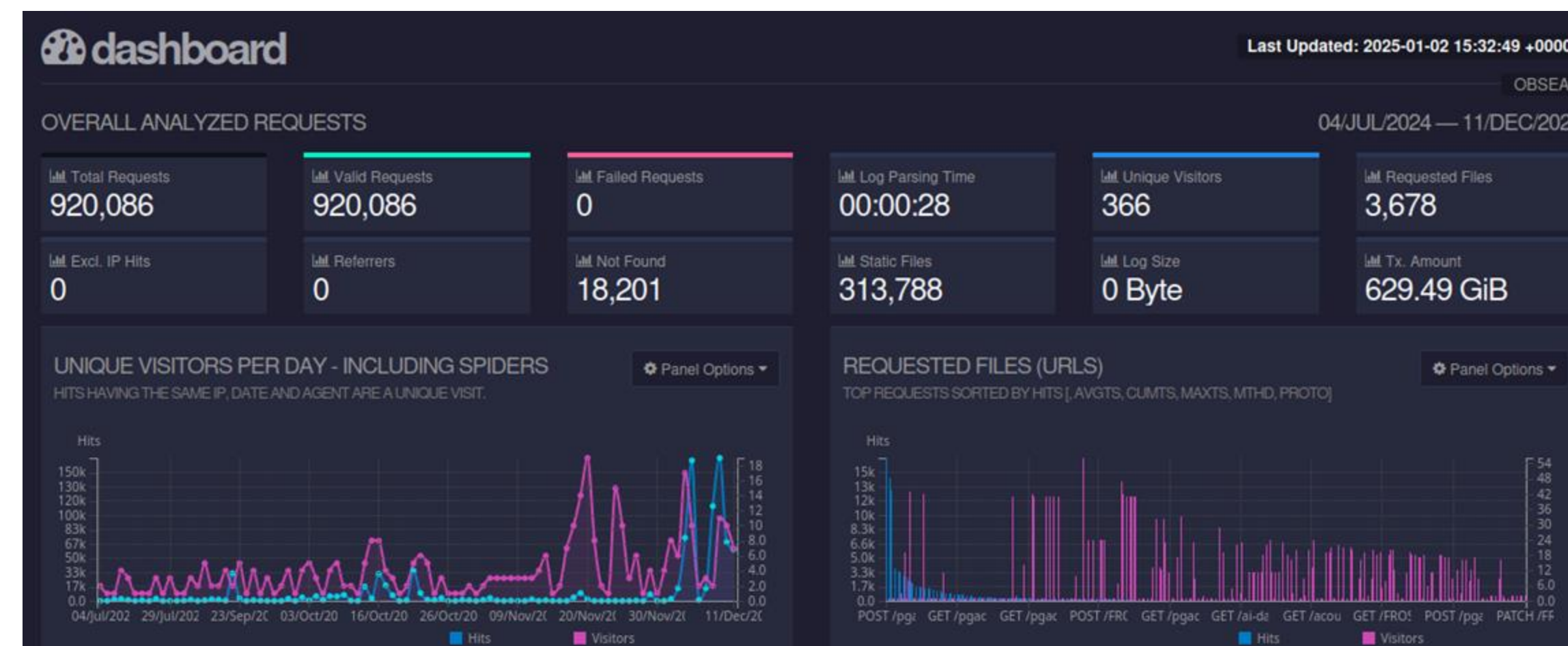
- Provides statistics on:
 - data access
 - number of downloads
 - number of users (IPs)
- Dependencies: ERDDAP

Priority: **HIGH**

Development Cost: 0.5 PM

Operational Cost: 1 PM

Target users: internal (RF leaders, data managers)



EMSO IT Services

Infrastructure Monitoring

- Monitors infrastructure performance
- Detects downtimes
- Trigger alarms
- **Dependencies:** ERDDAP

Priority: **HIGH**

Development Cost: 0.5 PM

Operational Cost: 1 PM

Target users: internal (data managers)

The screenshot displays the Zabbix Global view dashboard. The left sidebar shows navigation options: Monitoring (Dashboard, Problems, Hosts, Latest data, Maps, Discovery), Services, Inventory, Reports, Configuration, and Administration. The main content area shows 'Global view' with a summary of 'PRODUCTION Sensors Problems' (0 Disaster, 1 High, 19 Average, 0 Warning, 0 Information, 0 Not classified) and a table of 'PRODUCTION Problems'. A detailed alert for 'obsea_ci_alarms' is shown, indicating a high severity problem on host 'SBE37' starting at 13:52:56 on 2025.01.15. The alert details include the problem name, host, severity, and operational data.

Time	Info	Host	Problem • Severity	Duration	Ack	Action
23:18:04		egi-sta	vda: Disk read/write request responses are too high (read > 20 ms for 15m or write > 20 ms for 15m)	23m 59s	No	
Today						
2025-03-04 12:22:18		Vantage_Pro2	Vantage_Pro2_streams_no_data	7d 11h 19m	No	2
2025-03-04 11:22:31		vm1	/: Disk space is low (used > 80%)	7d 12h 19m	No	
2025-03-04 11:22:29		PostgreSQL vm1	PostgreSQL: Service is down	7d 12h 19m	No	2
2025						
2024-12-04 13:26:19		Grafana	is down	3M 7d 10h	No	2
2024-12-04 13:26:19		SensorThingsA	is down	3M 7d 10h	No	2

obsea_ci_alarms
❌ Problem: SBE37 SBE37_streams_no_data
Problem started at 13:52:56 on 2025.01.15
Problem name: SBE37_streams_no_data
Host: SBE37
Severity: High
Operational data: \$OBS01_CTD01,2025-01-15 11:52:34,2025-01-15 11:52:34,13.7825,4.19050,19.308,35.1605,1503.277
Original problem ID: 326671678 13:52

EMSO IT Services

Data Portal

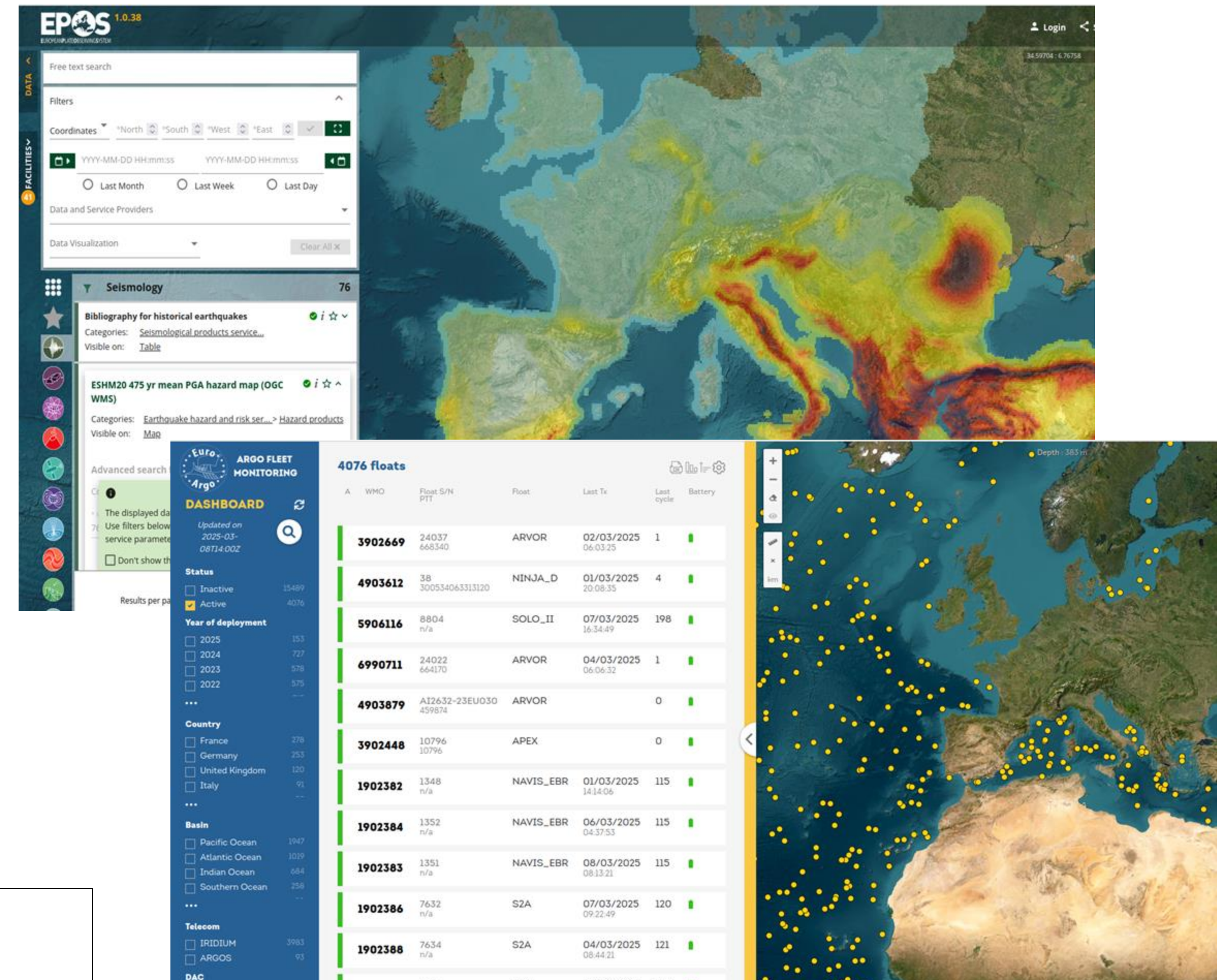
- User-friendly portal to access EMSO data
- Interactive Plotting
- Re-use existing solutions such as:
 - EURO ARGO
 - EPOS ERIC

Priority: **HIGH**

Development Cost: 4 PM

Operational Cost: 2 PM

Target users: external



EMSO IT Services

Connection with Data Aggregators

- Ensure that EMSO data is properly harvested and data is accessible through aggregators
- Technical tools (ERDDAP) already set up
- **Dependencies:** ERDDAP, Vocabulary

Priority: **HIGH**

Development Cost: 2 PM

Operational Cost: 1 PM

Target users: external



EMODnet



Copernicus
Marine Service



EMSO IT Services

Big Data Management

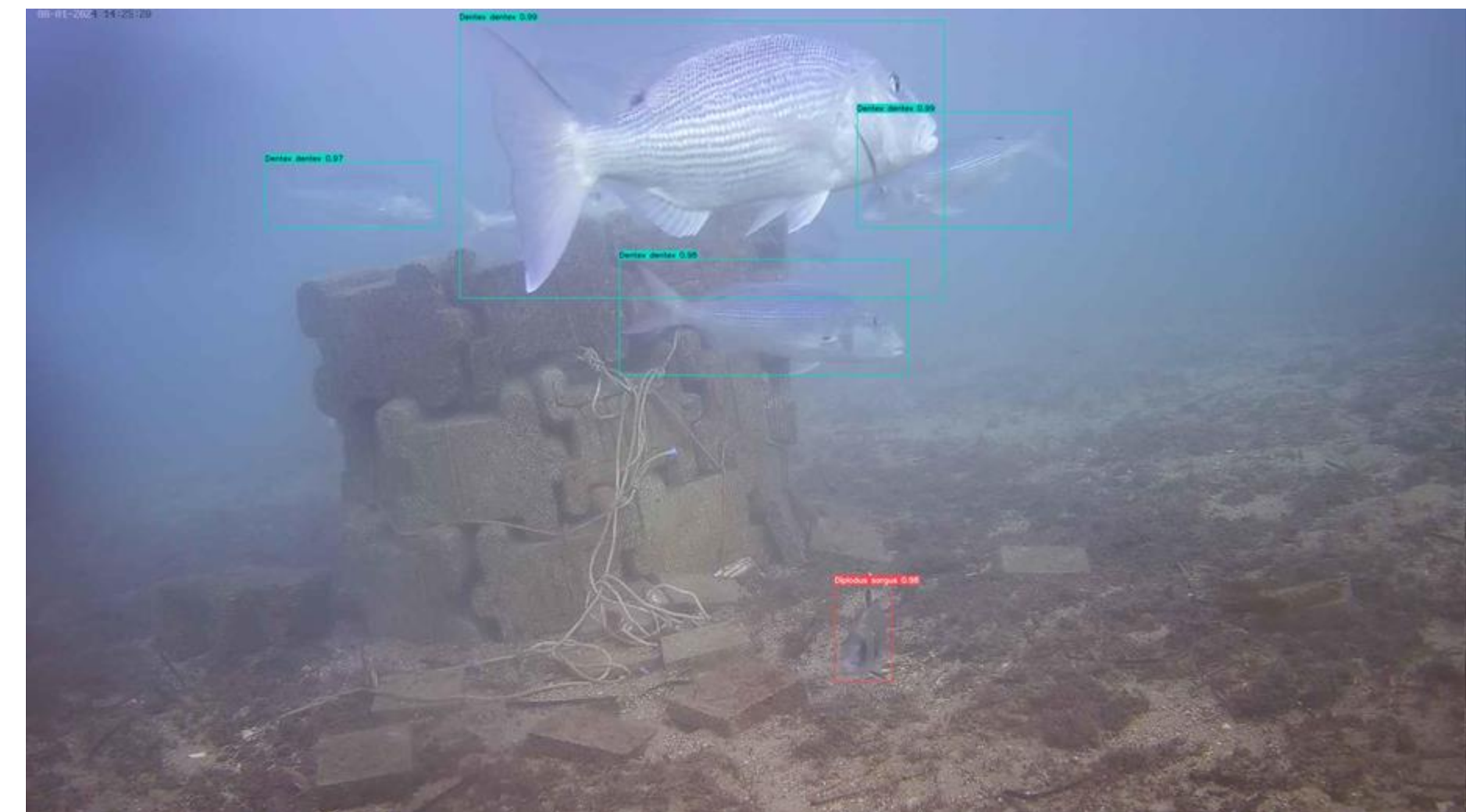
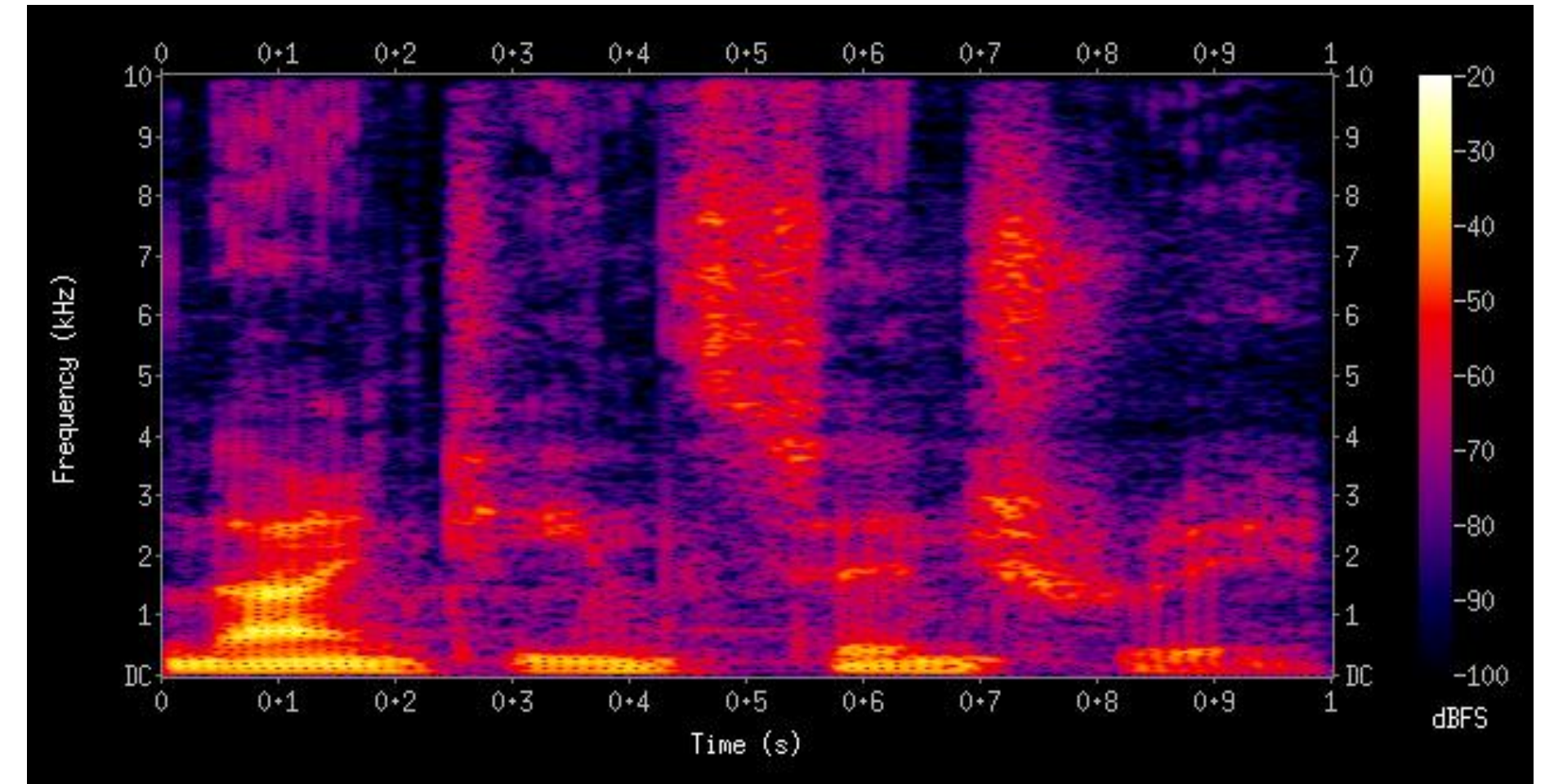
- Move beyond time series data
- Service to archive and serve file-based data
 - Acoustics, imagery...
- **Dependencies:** Vocabulary
- ***Added Value!***

Priority: **MEDIUM**

Development Cost: 10 PM

Operational Cost: 4 PM

Target users: internal / external



EMSO IT Services

Data Catalogue

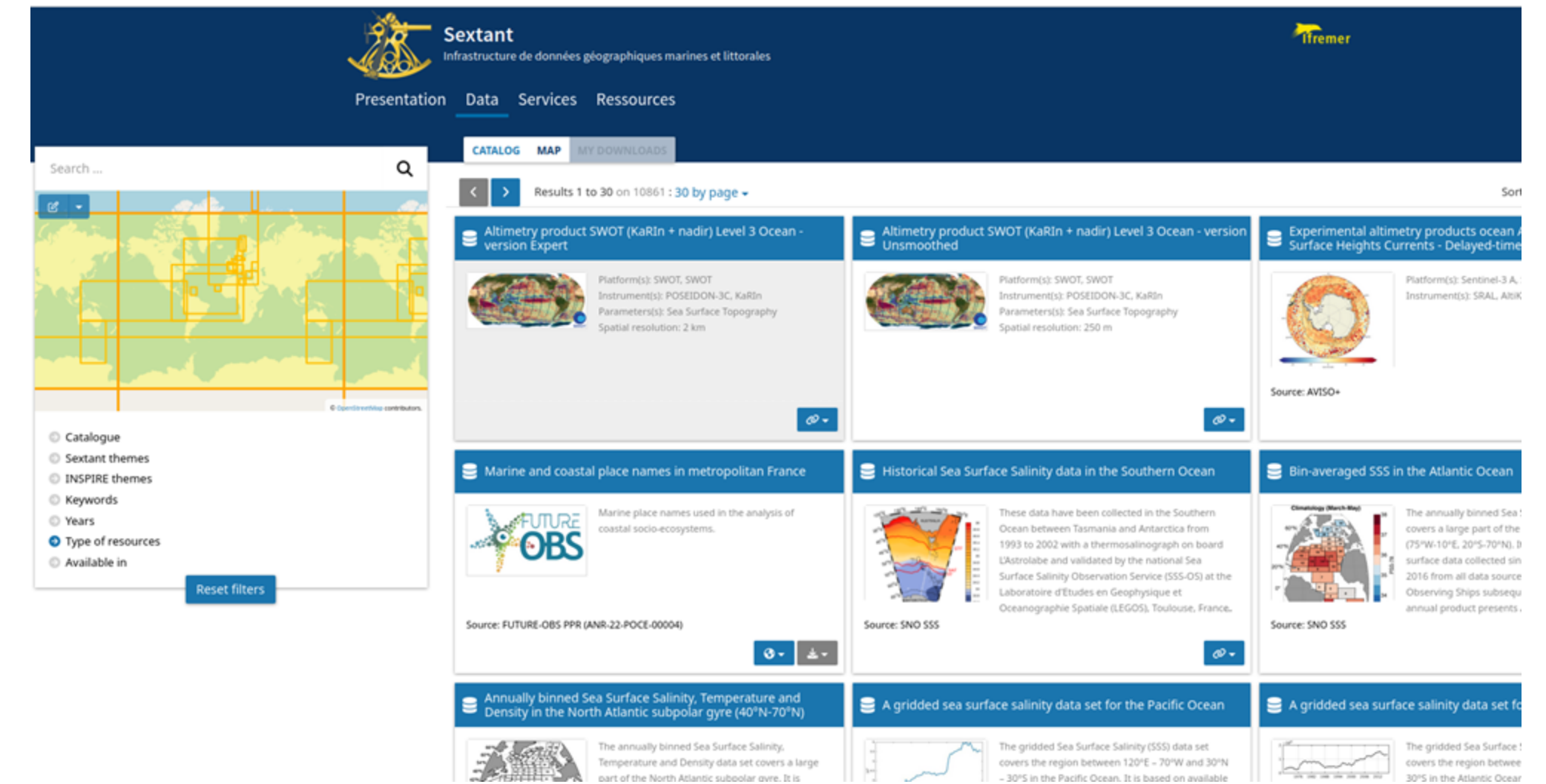
- Centralized catalogue that points to EMSO data
- Curated and well archived metadata
- Can be used to automatically mint DOIs
- **Dependencies:** Vocabulary, ERDDAP

Priority: **MEDIUM**

Development Cost: 3 PM

Operational Cost: 1 PM

Target users: internal / external



EMSO IT Services

Citation Reports

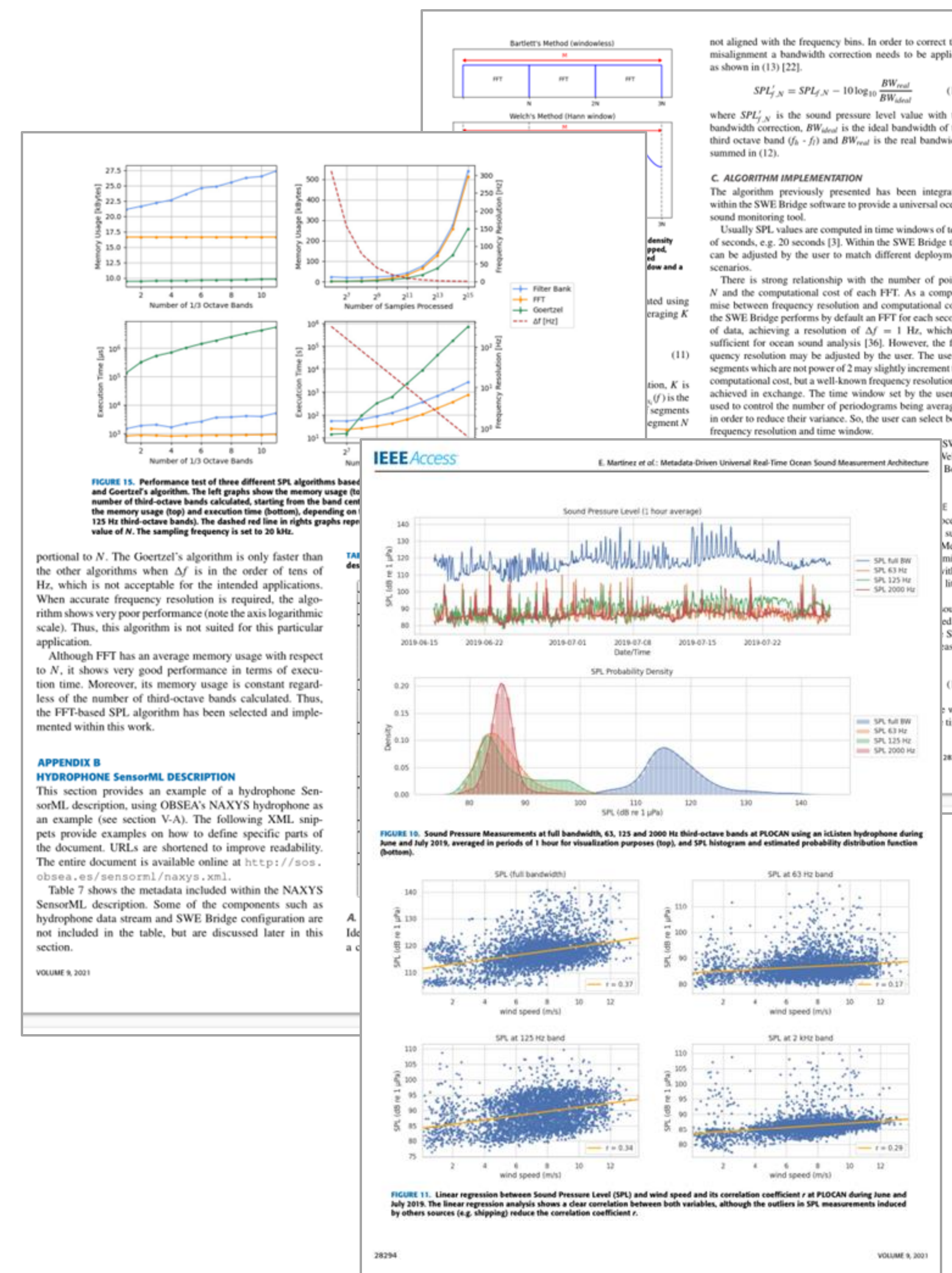
- Track usage of EMSO data across scientific papers
- Monitors scientific impact of EMSO
- Reporting needs

Priority: **MEDIUM**

Development Cost: 3 PM

Operational Cost: 2 PM

Target users: internal



EMSO IT Services

Quality Control / Quality Assurance

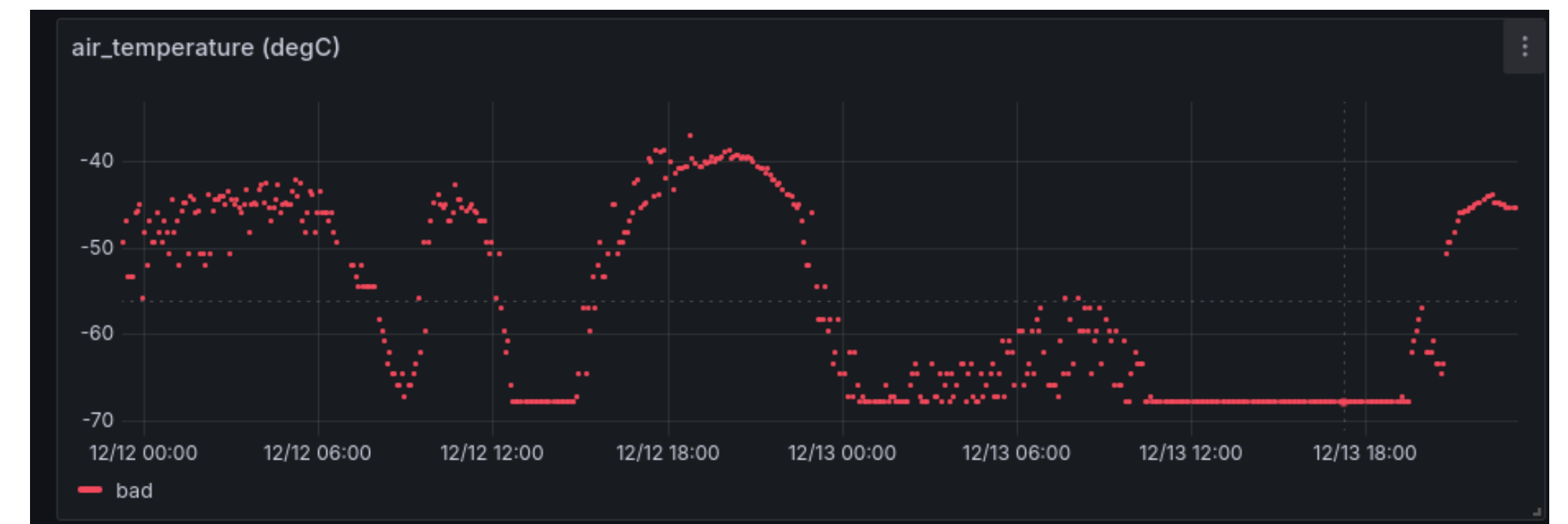
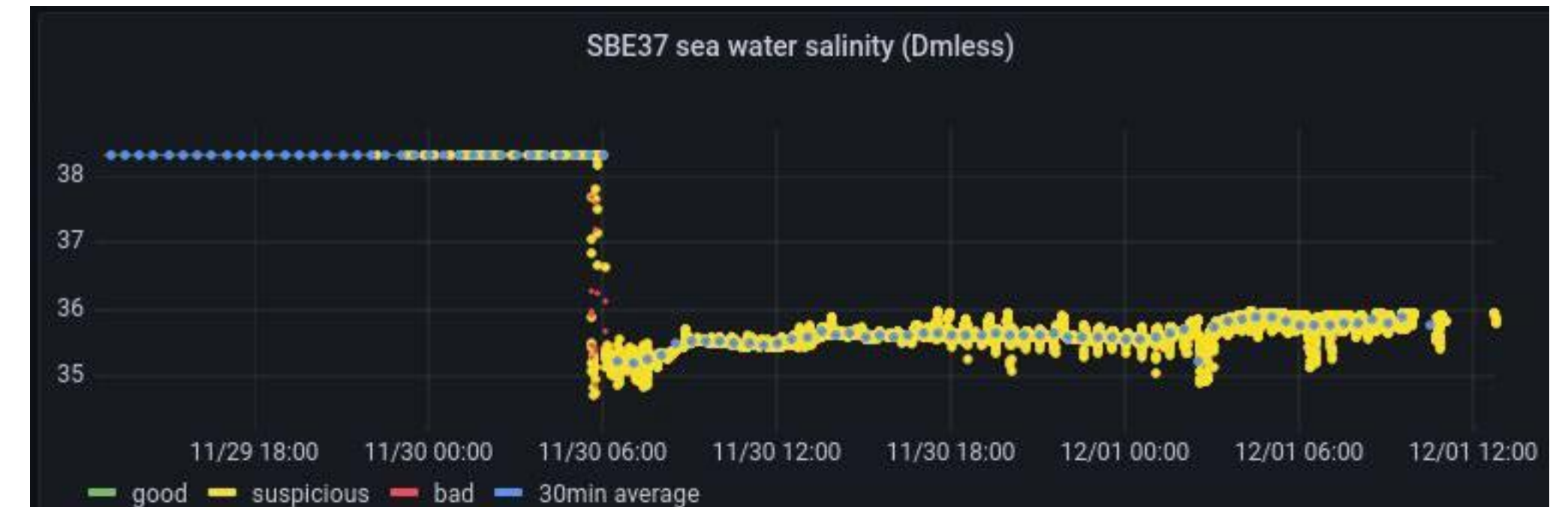
- Currently harmonized flagging scheme...
- ... but not harmonized protocols
- We need a common QC/QA methodology
- ⚠ It could disrupt existing workflows

Priority: **MEDIUM**

Development Cost: 3 PM

Operational Cost: 2 PM

Target users: internal



EMSO IT Services

Virtual Research Environment

- Notebook service
- Computational resources close to the data
- Very important for Acoustics, DAS, etc.

Priority: **LOW**

Development Cost: 3 PM

Operational Cost: 2 PM

Target users: internal / external



EMSO IT Services

Data Analysis

- Standardized data processing applied to ALL data:
 - e.g. Long-term trends of EOVs
- From simple calculations to complex AI workflows
- **Dependencies:** Virtual Research Environment (VRE)
- ***Added Value!***

Priority: **LOW**

Development Cost: 6 PM

Operational Cost: 2 PM

Target users: internal / external

EMSO IT Services

Uncertainty/Metrological Metadata

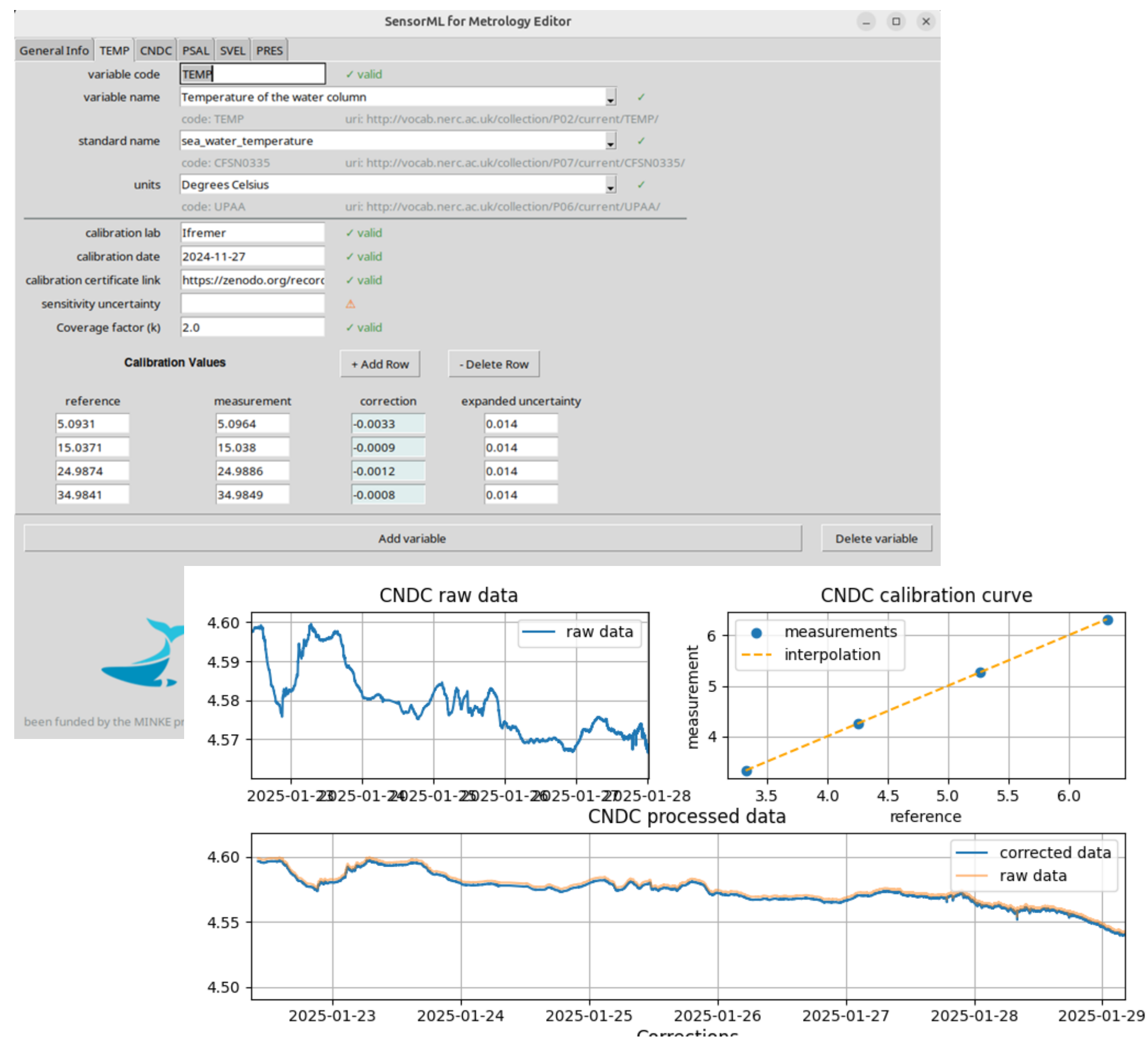
- Include Metrological metadata
 - Calibration certificates / Uncertainties
- Automated tools to correct data
- Legacy of MINKE EU project
- **Added Value!**

Priority: **LOW**

Development Cost: 6 PM

Operational Cost: 2 PM

Target users: internal / external

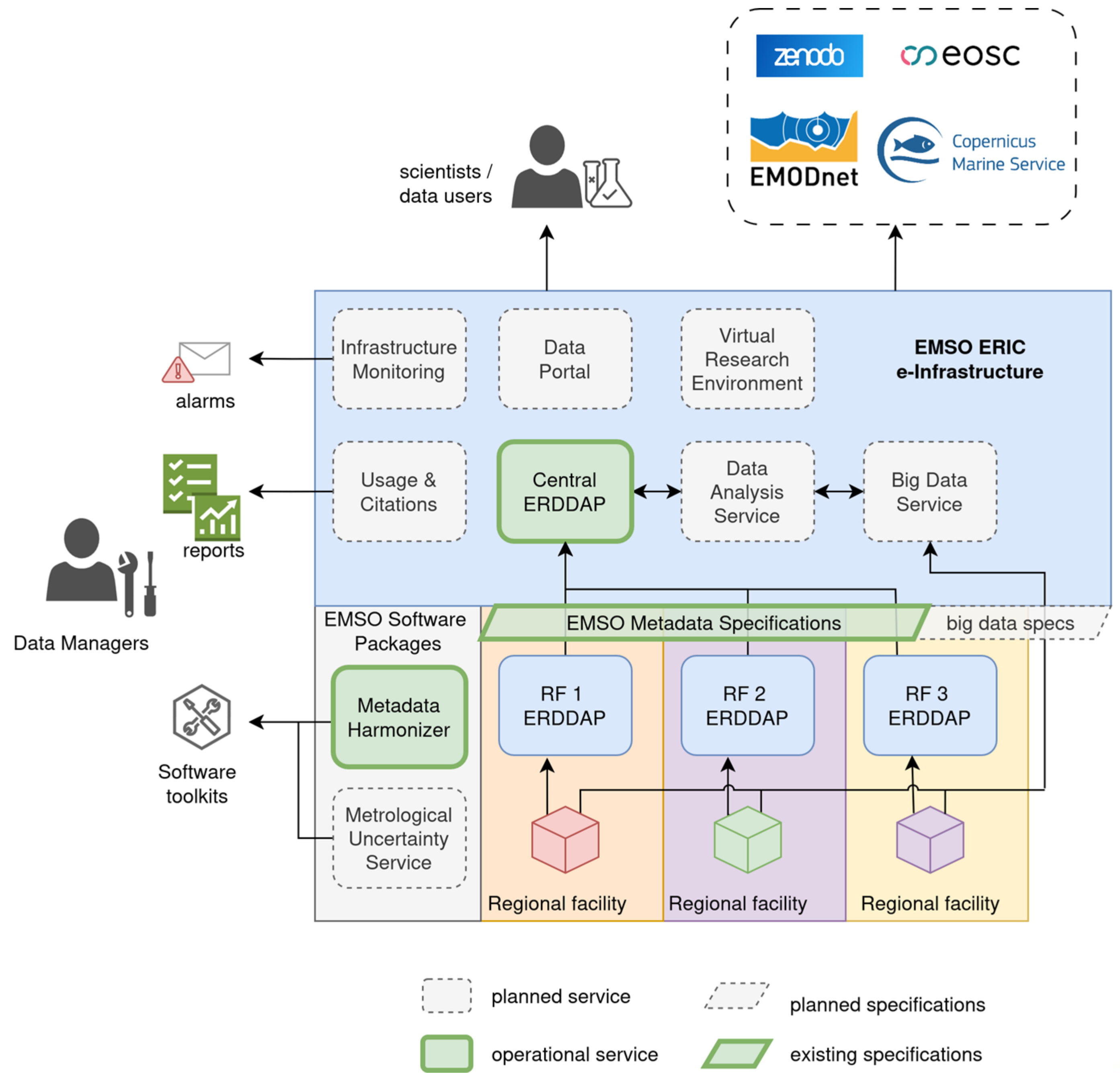


Conclusions

- Basic building blocks in place...
- ...but we need to move forward!
- Several new services to be implemented
 - > Foster new science
- In-kind contributions not enough
 - Funding required
 - Workload to be distributed among RFs

IT Service	priority	# Development PM (start-up cost)	# Operational PM (recurrent cost)
GoAccess Usage Statistics	HIGH	0.5	1
Infrastructure Health Monitoring	HIGH	0.5	1
Data Portal	HIGH	4	2
Connection with data aggregators	HIGH	2	1
EMSO Vocabulary	HIGH	1	0.5
Data Catalog	MEDIUM	3	1
Service for Big Data management	MEDIUM	10	4
Quality Control / Quality Assurance	MEDIUM	6	3
Data Repository Connector	MEDIUM	2	1
Automated citation report for data usage	MEDIUM	3	2
EMSO Virtual Research Environment (VRE)	LOW	6	3
Data analysis (including AI)	LOW	6	2
Inclusion of Calibration/Uncertainties in metadata	LOW	6	2
TOTAL		44	21.5

Thank you for your attention!





Observing the ocean to save the earth