EUROPEAN
PLASMA RESEARCH
ACCELERATOR WITH
EXCELLENCE IN
APPLICATIONS

EuPRAXIA
Preparatory Phase

WP7 – e-needs and Data Policy

Ricardo Fonseca (IST/Portugal) for WP7

Funded by
the European Union

This project has received funding from the European Union´s Horizon Europe research and innovation programme under grant agreement No. 101079773. It is supported by in-kind contributions by its partners and by additional funding from UK and Switzerland.

WP7 goals and deliverables

Funded by
the European Union

Workpackage goals

- **Define e-needs and data policy**
- **Connect to ongoing integrating projects in EU**
- **Establish a conceptual design of the e-infrastructure**

Benchmark of e-needs of similar facilities and infrastructure in Europe

MS7.1

Benchmark within existing facilities

Survey of e-needs and data policy at operating photon / neutron / light-source facilities offering comparable capabilities.

D7.1

Data Management Plan

First DMP for the operational phase, prepared to the Horizon Europe DMP template v1.1 (Art. 17, Open Science).

D7.2

Integration into the European landscape

Positioning against EOSC, EuroHPC JU and GÉANT; proposal of the five-line EuPRAXIA data strategy.

D7.3

Final report - e-needs and data policy

Data policy proposal, updated DMP, and the conceptual design of the e-infrastructure. The subject of this talk.

A data-intensive facility under an Open Science mandate

Funded by
the European Union

~100 PB

generated per year

at full operation, mostly for internal use

10-20 GB/s

peak, full experiment

1-10 GB/s per instrument during a run

2-4 PB

published per year

the subset selected for preservation

≥ 10 years

retention

open, FAIR, with metadata kept indefinitely

Two commitments that have to be met simultaneously

- Serve users in ultra-fast science – high-resolution medical imaging, positron annihilation spectroscopy, research on biomolecules and viruses – from two plasma-accelerator sites, INFN-LNF (beam-driven) and ELI-Beamlines (laser-driven).
- Comply fully with the European Commission Open-Science policy and FAIR data principles for all EC-funded work.

Challenge goes beyond data storage and compute

- Focus on policy, curation and governance
- In place for the start of operational phase in ~2028

WP7 contribution to EuPRAXIA in 5 parts

Funded by
the European Union

1

Data strategy and data policy

Five strategy lines; a policy built on the PaNOSC FAIR framework

2

Data management and the DMP

Automated metadata, PIDs, quality assurance, updated DMP

3

e-infrastructure conceptual design

Three-tier architecture, workflows, European integration

4

Requirements and resource needs

Functional / non-functional requirements; sizing estimates

5

Risks and open decisions

What must be resolved before technical design

Data strategy and data policy

Funded by the European Union

1

Data strategy and data policy

Five strategy lines; a policy built on the PaNOSC FAIR framework

2

Data management and the DMP

Automated metadata, PIDs, quality assurance, updated DMP

3

e-infrastructure conceptual design

Three-tier architecture, workflows, European integration

4

Requirements and resource needs

Functional / non-functional requirements; sizing estimates

5

Risks and open decisions

What must be resolved before technical design

The EuPRAXIA data strategy along 5 lines

Funded by the European Union

1. **Open-Science data policy**

   Enable open-science practice by users, leading to publication of data under FAIR principles.

2. **Data management infrastructure**

   Dedicated infrastructure for the collection and preparation of EuPRAXIA data for publication.

3. **Collaboration with a trusted repository**

   Inside the project or through EOSC services, for long-term storage and publication.

4. **External computing collaborations**

   Access to data centres and HPC resources for high-volume offline processing.

5. **Network interconnectivity**

   Adequate connectivity between EuPRAXIA sites, repositories, external services and users.

Guiding principle

Align with EU and EOSC recommendations wherever possible, and integrate with existing infrastructure rather than rebuild it.

Every line points outward: to EOSC for publication and metadata, to EuroHPC for compute, to GÉANT and the NRENs for connectivity, and to an established trusted repository for preservation.

EuPRAXIA should build what is specific to EuPRAXIA - acquisition, on-site processing, metadata capture - and buy or federate the rest.

Data policy based on PaNOSC FAIR framework

Funded by
the European Union

Why PaNOSC

- Common policy framework for the photon and neutron community
- Widely adopted: the basis of most published policies at comparable European facilities
- Comes with implementation notes on curation period, licences and persistent identifiers

Status and next step

The full policy text is Annex I of D7.3.
It has not yet been approved by the governing board.

The policy also constrains technology: format choices, catalogue design and the split of responsibilities between facility and user all follow from it. Any final decision on the policy has to be taken with those consequences in view.

Key points of the proposed policy

- **Curation** The facility is the curator: it is responsible for storing and curating the data;
- **Openness** Data becomes openly available after an embargo period, under an appropriate license;
- **Findability** The facility publishes metadata in a searchable online catalogue linking metadata to raw data;
- **Retention** Data preserved for a minimum of ten years; metadata stored indefinitely;
- **Readability** Well-defined formats, with the tools needed to read the data made publicly available;
- **User duty** Users are responsible for supplying accurate contextual metadata so that data meets FAIR principles.

Specific to EuPRAXIA

Funded by
the European Union

Processed data is not curated long-term

- Long-term curation focuses on raw data plus metadata
- Exception: a subset of user-processed data, up to an agreed volume, may be curated on PI request - with metadata describing the provenance of the processing
- Auxiliary data is curated in its original format; readability is not guaranteed by the facility

Why, and what it costs us

- Storing raw plus processed data would improve reuse - end users would not need to reprocess - but increases storage cost
- Storing only processed data would cut cost but many processing steps are irreversible, so reuse potential is lost with the original
- A shorter retention for processed data (e.g. four years) is under consideration as a middle path

Governance that comes with the policy

Data management team

Hired from project funds, with defined roles - including responsibility for the quality assurance procedures.

Data management committee

Nominated to take data policy decisions, including which datasets and metadata are preserved.

Beam-time condition

Acceptance of the policy and of the DMP is a prerequisite for the award of beam time.

Data management and the DMP

Funded by the European Union

1

Data strategy and data policy

Five strategy lines; a policy built on the PaNOSC FAIR framework

2

Data management and the DMP

Automated metadata, PIDs, quality assurance, updated DMP

3

e-infrastructure conceptual design

Three-tier architecture, workflows, European integration

4

Requirements and resource needs

Functional / non-functional requirements; sizing estimates

5

Risks and open decisions

What must be resolved before technical design

FAIR compliance

Funded by
the European Union

1. **Persistent identifiers**
   PIDs generated automatically for all preserved data, for all instrumentation at both sites. PID policy follows the EOSC PID policy.

2. **Automatic metadata capture**
   Beam delivery and site parameters captured automatically; experiment description supplied by users from a standard template.

3. **Self-describing datasets**
   Instrument data automatically complemented by a README: instrument description, data structure, naming conventions, units.

4. **User metadata, with tooling**
   Users supply experiment-specific metadata - sample description and similar - through tools and workflows provided for the purpose.

5. **Transfer to the repository**
   Prepared data and metadata moved to a trusted repository using community-standard formats and metadata ontologies.

Design intent

Every step that can be automated is automated. What remains for the user is only what the facility cannot know - and that arrives through templates and workflow tools (not a blank form)

FAIR Compliance is the default, not extra work

Data quality assurance, along EOSC recommendations

Funded by
the European Union

Two quality dimensions

Fitness for use
The data is reliable and usable.

Fitness for purpose
The data carries the information users need to judge whether it fits their own requirements.

Reference
Procedures follow the recommendations of the EOSC FAIR Metrics and Data Quality Task Force.

Quality assurance is written into the DMP as a responsibility of the data management team - not left as good intentions.

Technical

Metadata and data file checks

Availability of metadata and documentation checked on upload, plus metadata compliance tests. Only data meeting the requirements is accepted.

Scientific

Content and cross-content checks

Regular soundness checks on site-generated data, cross-validated between instruments where possible. User data checks are the user's responsibility; suspect datasets are marked corrupt until rectified.

Stewardship

Accessibility and understandability

Yearly review of how well the technical checks actually deliver fitness for purpose, with metadata and documentation revised accordingly.

Data Management Plan for the operational phase

Funded by
the European Union

What it is

- Written to the Horizon Europe DMP template v1.1 (1 April 2022), following Article 17 (Open Science) of the General Model Grant Agreement
- An updated version of the D7.1 DMP; included as Annex II of D7.3
- A living document: reviewed at least every twelve months, and whenever a policy decision affecting data management is taken
- Acceptance is mandatory for beam-time access, defined per proposal and confirmed or updated up to the end of the embargo period

Baseline choices proposed in the DMP

Format	NeXus / HDF5 as baseline open, self-describing format
Metadata	NeXus metadata schema; no project-specific ontologies unless unavoidable, and published openly if created
Data licence	Creative Commons Attribution 4.0 (CC BY 4.0)
Metadata licence	Creative Commons Zero (CC0), harvestable by EOSC services
Embargo	Optional, up to three years from the end of the experimental campaign
Access model	Public users publish all data except for legal or ethical constraints; proprietary users decide, per beam-access contract

e-infrastructure conceptual design

Funded by the European Union

1

Data strategy and data policy

Five strategy lines; a policy built on the PaNOSC FAIR framework

2

Data management and the DMP

Automated metadata, PIDs, quality assurance, updated DMP

3

e-infrastructure conceptual design

Three-tier architecture, workflows, European integration

4

Requirements and resource needs

Functional / non-functional requirements; sizing estimates

5

Risks and open decisions

What must be resolved before technical design

Data characteristics and the run-based operating model

Funded by the European Union

High-rate instrument data

Dominated by images and numerical datasets. NeXus/HDF5 as the baseline open format. Acquired by facility instruments and DAQ, with provisions for specific user instrumentation.

Metadata

Beam parameters, facility conditions, instrument calibration, experimental configuration, project information - beyond the minimum needed for FAIR.

Auxiliary and processed data

Ancillary datasets where applicable, although long-term curation focuses on raw data plus metadata.

Run-based acquisition model

~1 HOUR

Data-taking run

Very high short-term rates: 1-10 GB/s per instrument, 10-20 GB/s for the full experiment, depending on beam repetition rate.

MINUTES

Transfer on site

On run completion, data moves to on-site intermediate storage for analysis, inspection and quality assurance.

CAMPAIGN END

Publication

Validated data plus metadata transferred to long-term storage and made publicly available under open data / FAIR requirements.

A three-tier storage and processing architecture

Funded by the European Union

I — High-performance acquisition

Buffers 30-60 min per instrument; detectors cabled at 10-40+ GbE or HDR InfiniBand.

- DAQ and fast buffer storage
- NVMe / SSD tier
- Online-processing workstation

1-10 GB/s per instrument

II — On-site processing and user area

Frees Level I for the next run, enriches metadata, and provides all user-facing access.

- Short-term storage + compute cluster
- Metadata catalogue and enrichment
- User portal and e-logbook

300-600 TB per day

III — Off-site trusted repository

Findable for ten years; third-party access over a free standardised protocol.

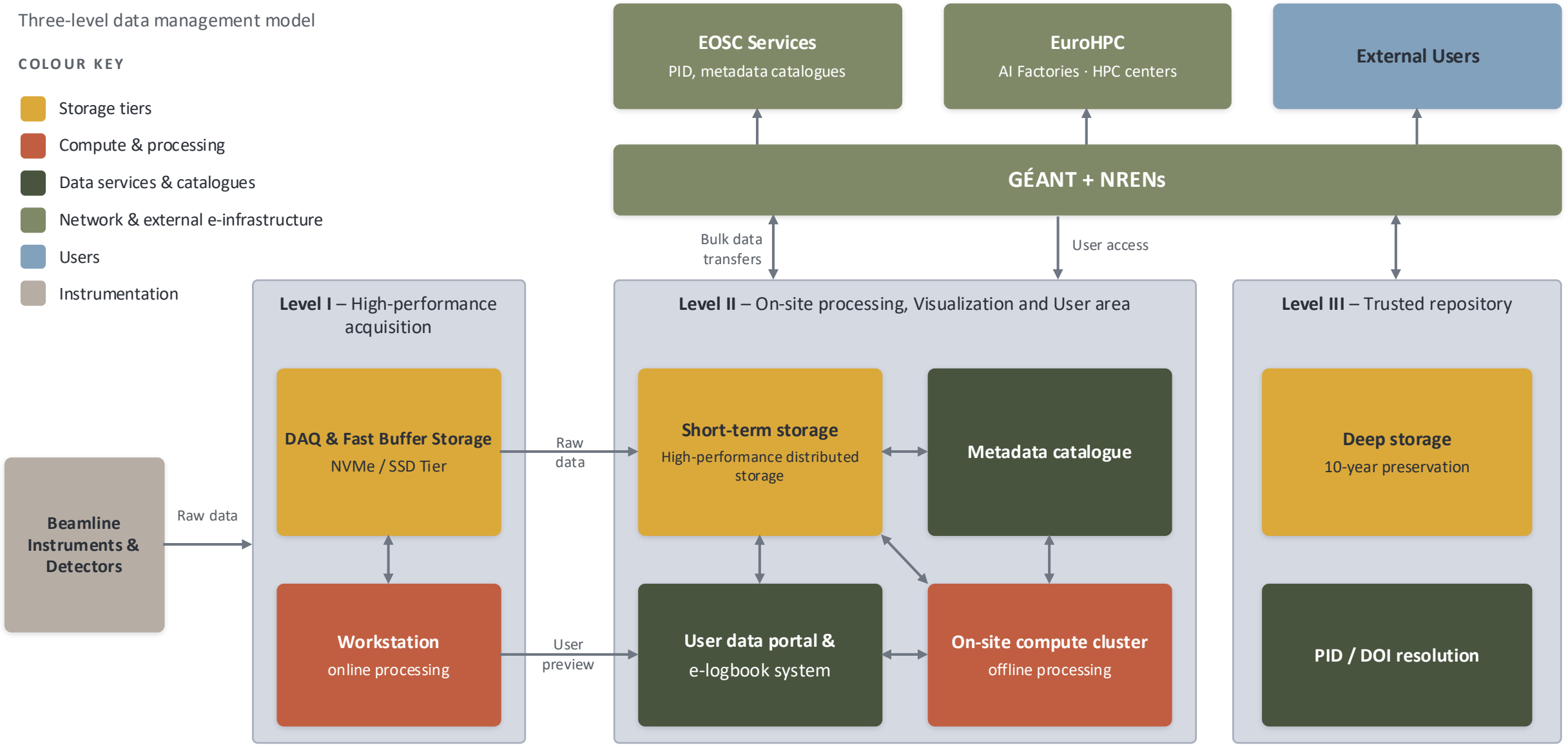
- Deep storage, 10-year preservation
- PID / DOI resolution
- EOSC-harvestable metadata

2-4 PB per year

Architecture overview

Funded by the European Union

Three-level data management model



Integration with European e-infrastructures

Funded by the European Union

EOSC

European Open Science Cloud

Federated environment for scientific data and services, promoting open science and FAIR principles.

EuPRAXIA will:

- Adhere to EOSC guidelines
- Provide metadata for harvesting
- Integrate the repository with EOSC nodes

EuroHPC JU

High-performance computing

State-of-the-art supercomputing and AI resources for European research and innovation.

EuPRAXIA will:

- Offload tomography and simulation
- Use AI Factories for ML-based reconstruction
- Batch and streaming workloads

GÉANT

Pan-European research network

High-speed connectivity for research and education communities, with the national NRENs.

EuPRAXIA will:

- Site-to-site and site-to-repository links
- External user access
- Additional bandwidth procured if HPC access demands it

Operational Workflows

Funded by
the European Union

Online

Real-time and near-real-time steps that occur during the data-taking run, enabling users to steer experiments based on immediate feedback

Offline

Processing and analysis that occur after the completion of a run, including quality assurance, enrichment with metadata, and automated pipelines

Publication

Transfer of curated, validated datasets to an external trusted repository for long-term preservation and FAIR-compliant open access

Design goal

Minimize user burden by automating, where possible, the data management process, from acquisition to publication in a trusted repository, while supporting both real-time decision-making and long-term FAIR compliance

Structure reflects EuPRAXIA's three-tier technical architecture as described in the previous slides

Online workflow: “real-time” feedback during a run

Funded by
the European Union

1 High-rate capture at Level I

Data written to high-speed storage directly attached to the acquisition hardware.

2 Preview data generation

Level I workstation performs rapid preprocessing – image correction, histogramming – and streams preview to e-logbook system.

3 Real-time monitoring

Users see online dashboards with plots, images and diagnostic metrics. Target latency: 1-10 seconds.

4 Automated metadata capture

Instrument state and facility parameters added automatically as the run proceeds.

5 Quality assurance

Technical checks on integrity and format compliance; scientific checks on detector sanity, beam quality and expected signal.

6 User steering

On the basis of online results, users may adjust beam, detector, timing or sample parameters.

Purpose: verify that detectors, instrumentation and beam are behaving, give the experimental team the feedback it needs to steer the experiment, and capture metadata while the context still exists.

Offline and publication workflows

Funded by the European Union

OFFLINE

After the run - minutes to hours per dataset

1. **Transfer Level I to Level II**
   Data moves from the fast buffers to larger on-site storage.
2. **Processing**
   Reconstruction, tomography, event filtering, machine-learning workflows.
3. **Optional off-site processing**
   Transfer to external HPC or AI systems, as batch or streaming jobs.
4. **Extended quality assurance**
   Technical, scientific and stewardship checks: file validity, metadata completeness, cross-instrument consistency.
5. **Metadata enrichment**
   Contextual metadata from users - sample, protocol - plus automatic injection.

PUBLICATION

After the campaign - custody moves off site

1. **Final validation and packaging**
   PIDs assigned; README files and workflow documentation included.
2. **Transfer to Level III**
   The trusted repository assumes custody of the data.
3. **Metadata publication**
   Exposed publicly under CC0, indexed and harvestable by EOSC services.
4. **Open data release**
   Accessible immediately unless a permitted embargo of up to three years applies.
5. **Long-term curation**
   The repository maintains integrity and availability over the retention period.

Responsibilities

Online: facility systems and experimental teams - acquisition, monitoring, metadata capture, steering; Offline: users and facility experts - processing, QA, metadata enrichment, run consolidation; **Publication**: data management team and repository - PID assignment, preservation, licensing, open access.

Requirements and resource needs

Funded by the European Union

1

Data strategy and data policy

Five strategy lines; a policy built on the PaNOSC FAIR framework

2

Data management and the DMP

Automated metadata, PIDs, quality assurance, updated DMP

3

e-infrastructure conceptual design

Three-tier architecture, workflows, European integration

4

Requirements and resource needs

Functional / non-functional requirements; sizing estimates

5

Risks and open decisions

What must be resolved before technical design

e-infrastructure requirements

Funded by the European Union

FUNCTIONAL - what the system must do

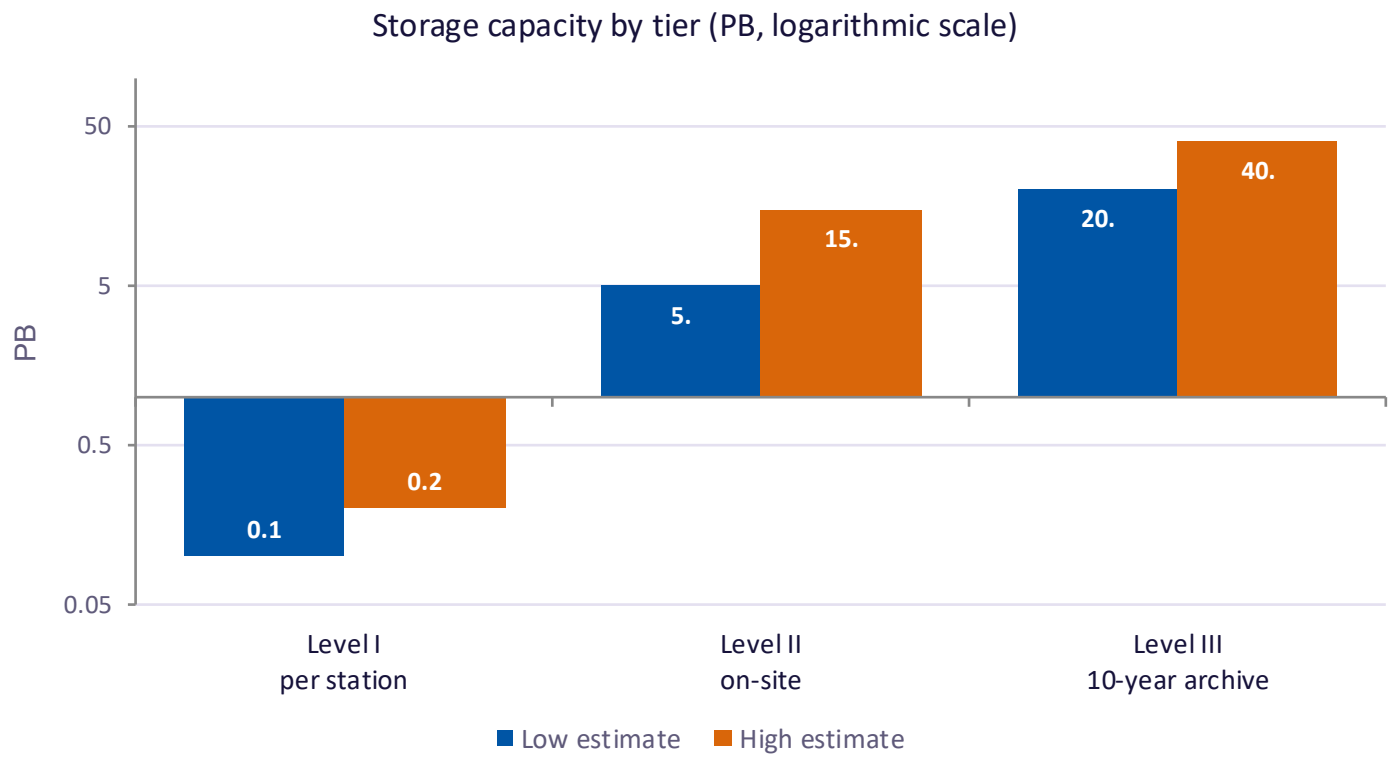
FR-1	Capture 1-10 GB/s per instrument, 10-20 GB/s per experiment, sustained through a run
FR-2	Automatic metadata capture: beam, facility state, instrument configuration, calibration, timing
FR-3	Real-time user feedback with latencies of a few seconds
FR-4	Online and offline processing pipelines, with offloading to external HPC
FR-5	Persistent identifiers for every dataset selected for preservation
FR-6	FAIR-compliant publication: appropriate licences, EOSC-harvestable metadata
FR-7	Differentiated access rights for staff, experimental teams and external users; optional embargo
FR-8	Long-term preservation of published raw data, ideally ten years or more, with metadata
FR-9	Seamless transfer of curated data to a trusted repository, internal or external

NON-FUNCTIONAL - how it must behave

NFR-1	Reliability: the acquisition toolchain must not lose data under any expected condition
NFR-2	Scalability: with repetition rate, additional beamlines, higher-resolution detectors
NFR-3	Interoperability: formats and schemas aligned with community standards
NFR-4	Performance: ≥10 GbE internal baseline, expandable to 40-100 GbE; 1-10 PFLOPs offline
NFR-5	Usability: intuitive portal, e-logbook and visualisation; workload reduced by automation
NFR-6	Security: authentication, role-based access control, controlled visibility of user data
NFR-7	Sustainability: media lifecycle, energy, and facility/user cost-sharing options
NFR-8	Preservation integrity: checksum verification and periodic audits over the retention period
NFR-9	Compliance with open-science regulation and the EuPRAXIA data policy

Estimated storage needs across the three tiers

Funded by the European Union



Level I

A 1-hour run at 5 GB/s yields ~18 TB per instrument, so 36-72 TB for a full setup. Recommendation: ≥100-200 TB of NVMe or equivalent per experimental station.

Level II

At eight runs per day, 300-600 TB/day accumulates during a campaign. Recommendation: 5-15 PB of high-performance distributed storage.

Level III

Only a subset of raw data is preserved. Benchmarking against other light sources suggests 2-4 PB/year, so 20-40 PB for a ten-year retention.

Design parameters - repetition rate, detector configuration, number of operational beamlines - are not yet fixed, so these estimates rest on the preparatory-phase benchmark values.

Compute and network provisioning

Funded by the European Union

COMPUTE

200 TFLOPs

per instrument, online

Memory-rich, GPU-accelerated workstations co-located with the fast storage, for correction, low-latency preview generation and QA.

5-10 PFLOPs

on site, offline

For tomography, reconstruction, machine learning and simulation. High-volume batch work offloaded to external HPC.

NETWORK

Detector → Level I	≥ 40-100 GbE or HDR InfiniBand
Level I → Level II	≥ 10 GbE; 40 GbE recommended for future-proofing
Cluster ↔ Level II	100 GbE or InfiniBand
Site uplink	10-100 GbE, within GÉANT/NREN capability

How this is meant to be funded

Eligible costs

Storage, metadata generation and personnel - including the data management team - covered by future Horizon Europe grants.

In-kind contributions

Long-term storage and access costs may be partly covered in kind by the associated open-data repositories.

Cost sharing with users

Model under exploration: EuPRAXIA supports site-generated data, users support end-station data and any storage beyond the standard volume or period.

Risks and open decisions

Funded by the European Union

1

Data strategy and data policy

Five strategy lines; a policy built on the PaNOSC FAIR framework

2

Data management and the DMP

Automated metadata, PIDs, quality assurance, updated DMP

3

e-infrastructure conceptual design

Three-tier architecture, workflows, European integration

4

Requirements and resource needs

Functional / non-functional requirements; sizing estimates

5

Risks and open decisions

What must be resolved before technical design

Six risks to carry into the next EuPRAXIA phases

Funded by the European Union

Structural

1 Facility parameters still evolving

Beam rate, beamline count and detector specifications directly set data rates, storage volumes and compute needs.

Risk: Incorrect sizing: bottlenecks or overspending.

2 Repository selection

Internal repository gives tight integration at higher effort; external reduces effort but creates dependency.

Risk: Delayed or suboptimal integration with preservation and publication.

3 Long-term storage funding

No funding model yet identified for ten-year curation; may need partial user support.

Risk: Insufficient budget, and a mismatch with user expectations.

Practice

4 Data reduction policy

Reduction techniques could offset storage needs, but risk removing relevant data or introducing bias.

Risk: Loss of scientific content, harming reproducibility and FAIR compliance.

5 Data quality and metadata

FAIR compliance depends on user-supplied contextual metadata; practice varies widely between communities.

Risk: Poor metadata reduces reusability and compromises FAIR principles.

6 Organisation and staffing

A data management team and a data management committee both still need to be recruited and nominated.

Risk: Under-compliance with open-data principles and delayed publication.

Four decisions EuPRAXIA will need to make

Funded by
the European Union

1	**Final data policy scope**	The proposed policy covers most issues, but two elements remain open: how data is selected for publication, and where the boundary lies between user and facility responsibility.
2	**Long-term repository strategy**	Build an internal repository, or adopt an external or partner repository. This drives cost, workflow and, above all, long-term sustainability.
3	**Standardisation and data reduction**	NeXus/HDF5 is the baseline, but final instrument designs may require additional schemas. Separately: whether to reduce data at all, and how to validate it.
4	**User access and funding model**	Embargo terms, access for proprietary versus public users, and how storage costs are divided between facility and users.

These must be resolved before finalizing the operational design: each one changes hardware sizing, workflow design or cost allocation.

Funded by
the European Union

Foundations for the EuPRAXIA data ecosystem

Policy — A data policy rooted in the PaNOSC FAIR framework: a governance model for ownership, curation, publication and stewardship, with responsibilities defined for facilities and users.

Plan — An updated Data Management Plan to the Horizon Europe template, with quality assurance procedures and acceptance as a condition of beam time.

Design — A three-tier conceptual architecture - acquisition, on-site processing, long-term preservation - supporting real-time steering, advanced analysis and FAIR publication.

Integration — Alignment with EOSC, EuroHPC and GÉANT, so that EuPRAXIA joins the European data landscape rather than duplicating it.

Thank you!