The Extreme Light Infrastructure
EXTREME SCIENCE

ELI FAIR High-Power Laser Experiments

Teodor Ivanoaica

Sr. Coordinator for Scientific Computation
and Data Management

A European Research Infrastructure Consortium

ELI ERIC is a single, multi-site organisation

A European Research Infrastructure Consortium – an ERIC

The Czech Republic, *Founding Member*
Host of Seat

Hungary, *Founding Member*
Host

Italian Republic
Founding Member

Lithuania
Founding Member

Federal Republic of Germany
Founding Observer

Bulgaria
Founding Observer

ELI ERIC involves the Czech Republic, Hungary, Italy and Lithuania as founding Members. Both Germany and Bulgaria are Founding Observers. ***Romania and ELI-NP are also expected to join the ELI ERIC consortium, which is open to European and non-European countries to join its membership.***

))))) eli

The mission of ELI ERIC is to provide access for European and international researchers to the ELI Facilities in the Czech Republic and Hungary.

ELI Attosecond Light Pulse Source
www.eli-alps.hu

ELI Beamlines
www.eli-beams.eu

The ELI ERIC Facilities

- **Data Challenges**
- **Data Challenge and FAIR Data Policy**
- **Data Management Plan approach**
- **FAIR Data @ ELI DOI's**
- **PaNOSC services @ELI**

- **Objectives and ELI's mission**
- **Data Management Plan, a multi-stage approach**
- **ELI Data Structure in the context of the 1st ELI ERIC Call**
- **ELI Users' Journey from Proposal to Data Access**

ELI Attosecond Light Pulse Source
www.eli-alps.hu

ELI Beamlines
www.eli-beams.eu
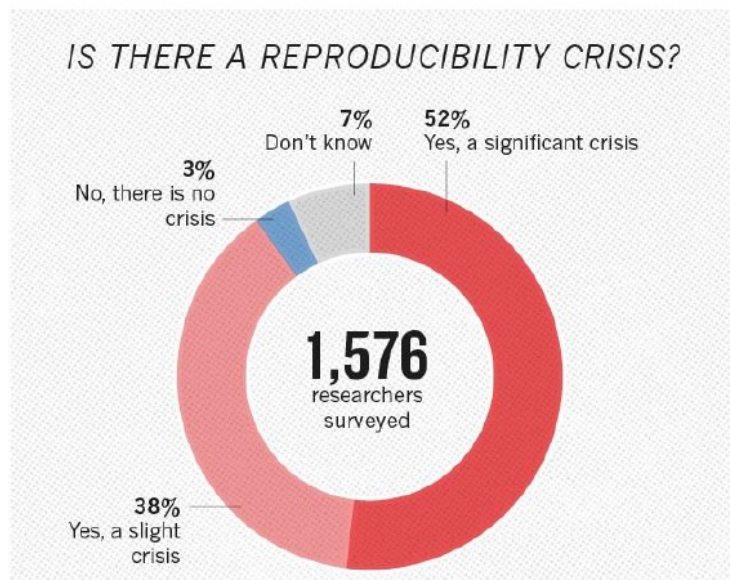
Access Modes to ELI ERIC

- **Excellence-Based Access** – Scientific evaluation of proposals by and international peer-review panels composed of qualified scientists. Results of experiments based on excellence must be published and open.
- **Mission-Based Access** – Thematic areas of research granted on the basis of specific scientific missions pursuing clearly defined challenges. Results of experiments generally published and open.
- **Proprietary Access** – Paid access for industrial or other users, where results are retained by the user, consistent with ELI ERIC's Data and IPR Policy.
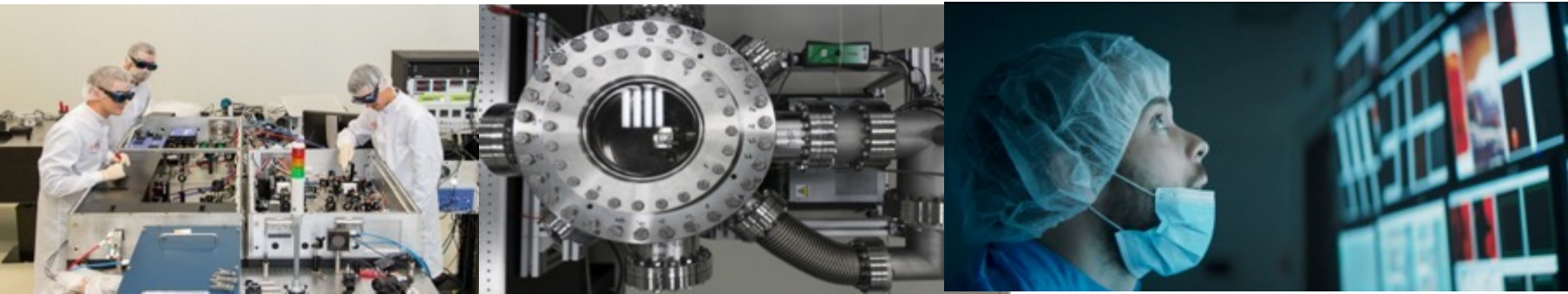
Is there a Reproducibility Crisis? What are the causes? How can it be addressed?

Nature **533**, 452–454 (2016) | Cite this article
5320 Accesses | **1225** Citations | **3871** Altmetric | Metrics

IS THERE A REPRODUCIBILITY CRISIS?

7% Don't know
52% Yes, a significant crisis
3% No, there is no crisis
**1,576** researchers surveyed
38% Yes, a slight crisis

- *A lack of access to methodological details,* **raw data***, and research materials*

- *Use of* **misidentified***, cross-contaminated, or over-passaged cell lines and microorganisms.*

- **Inability to manage complex datasets**

- **Poor research practices and experimental design**

- *A competitive culture that rewards novel findings and undervalues* **negative results (But are there really NEGATIVE Results or is it just a different perspective?)**

1. https://en.wikipedia.org/wiki/Replication_crisis
2. https://phys.org/news/2017-03-science-crisis.html

The ELI Context
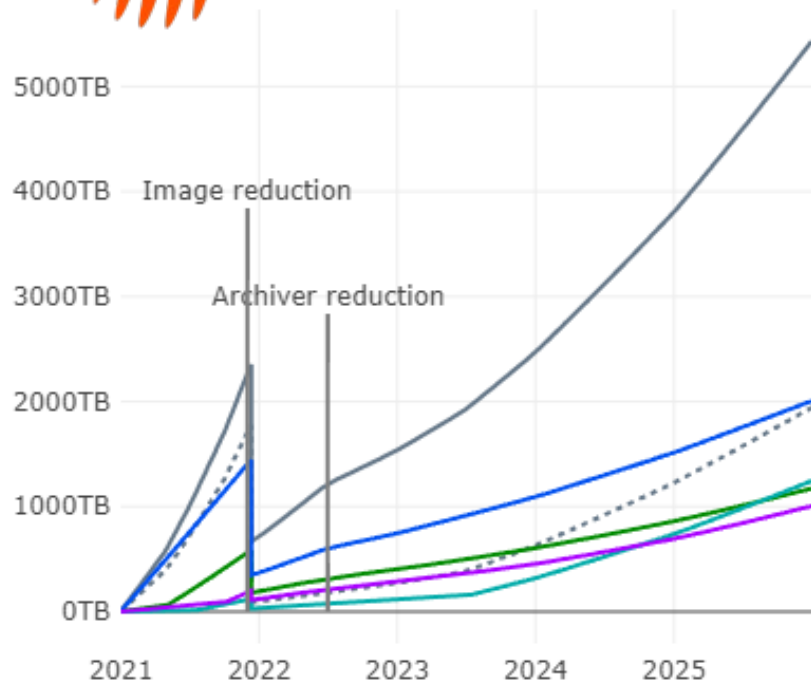
Total Production
Images
L1
L2
L3
I 4

Policy&DMP provide quality

Data Quality becomes crucial
Data is not so "easily searchable"
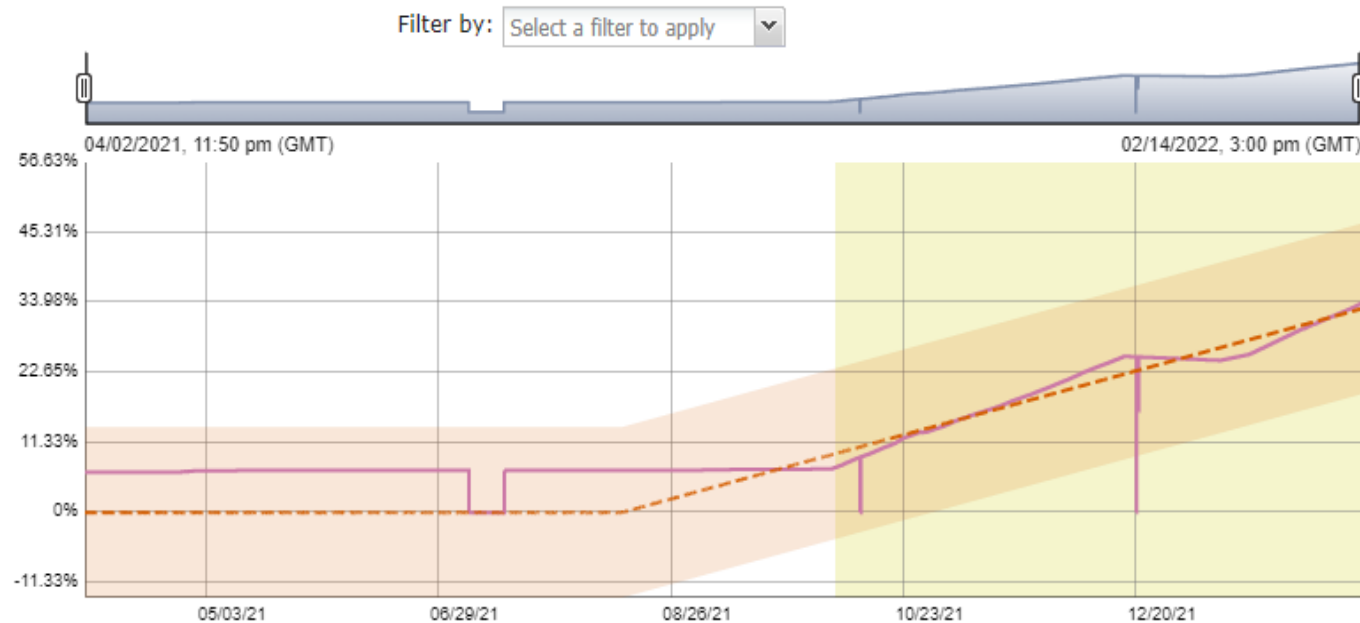Data becomes (not so) findable

Capacity Forecast

Filter by: Select a filter to apply

Plot data:
- Total Capacity
- Provisioned Capacity
- Writeable Capacity
- 90% of Provisioned
- Total Usage

Forecast data:
- Calculation Range
- Forecast Usage
- Standard Deviation
- Outliers

Y-axis units:
- Percent
- Bytes

04/02/2021, 11:50 pm (GMT)   02/14/2022, 3:00 pm (GMT)

Double-click the Forecast chart to clear the Calculation Range.

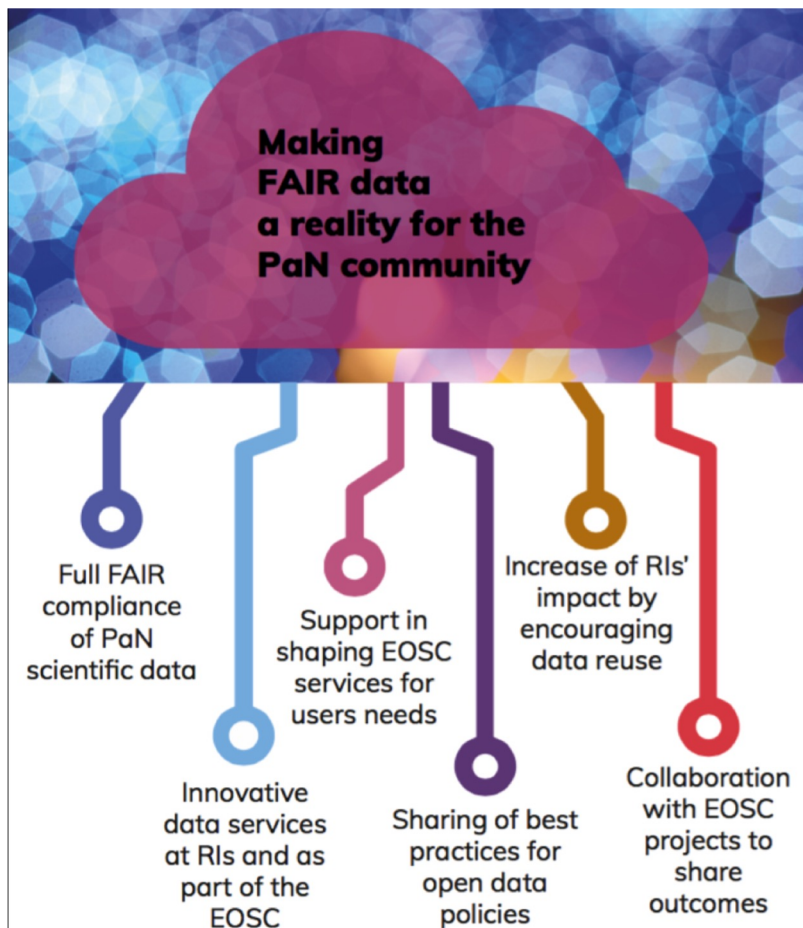You will reach **90% capacity** in **112 days** on Dec 30 2022.   Click to view

Capacity growth of 20 TiB per week is based on the calculation range.
The linear regression function has an R-value of 0.98.

What PaNOSC provides!

FAIR Data Policy framework and guidelines
Data Stewardship

Common domain specific search API / Open Data catalogues E-Logbook
Data Catalogue

Umbrella AAI solution based on eduTEAMS from GÉANT
EOSC Integration

Standardising metadata through the Nexus metadata ontology
Data Management

Simulation services for experiments and optical elements at PaN beamlines McStasScript python API for McStas
Data simulation

HDF5 web viewer HDF5 Jupyterlab viewer
Data Analysis

Remote data analysis portal - VISA
Data Analysis

Reproducible data analysis with Jupyter notebooks Jupyter slurm spawner
Data Analysis

Training portal
E-learning

Multi-disciplinary   PaNOSC

panosc
photon and neutron open science cloud

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852.

Making
FAIR data
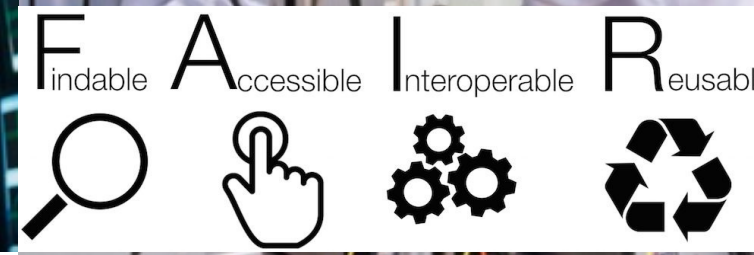a reality for the
PaN community

Full FAIR compliance of PaN scientific data

Innovative data services at RIs and as part of the EOSC

Support in shaping EOSC services for users needs

Sharing of best practices for open data policies

Increase of RIs' impact by encouraging data reuse

Collaboration with EOSC projects to share outcomes

10 Primary outcomes [of 42 in total]

1. FAIR data policy and DMPs
2. (FAIR assessment and common PID framework)
3. Standardised metadata (Nexus/HDF5, PaN ontologies)
4. Federated search API for PaN data catalogues
5. Open Data portal for searching + downloading data
6. Community AAI UmbrellaId
7. JupyterLab notebooks and HDF5/NeXus files visualisation
8. Remote data analysis with VISA + interoperable pipelines
9. Simulation software for simulating experimental data (SIMEX)
10. PaN-learning platform (pan-learning.org + pan-training.org)

3

ELI ERIC FAIR Data Policy

10.5281/zenodo.6515903

⌄	**Providing policy context**
⌄	**Defining research data**
⌄	**Policy registration**
⌄	**Explaining FAIR**
⌄	**Clarity on data sharing expectations**
⌄	**Making exceptions to data sharing clear**
⌄	**Updating DMPs**
⌄	**Clear roles and responsibilities**
⌄	**Clarity on eligibility of RDM costs**
⌄	**Help to find data repositories**

Table of Content

2. Recommendations
3. Generic scientific data management policy
   3.1. Definitions
   3.2. General principles
   3.3. Persistent identifiers
   3.4. Raw data and associated metadata
   3.5. Processed data and associated metadata
   3.6. Auxiliary data
   3.7. Results
4. Good practices
5. Termination of custodianship
APPENDIX 1 - Implementation Notes (IN)
APPENDIX 2 - FAIR Data Maturity Model analysis

Findable Accessible Interoperable Reusable

ELI ERIC FAIR Data Policy

- Prepare data for the exclusive use of the scientists who conducted the experiment which produced the data for up to three years after the conclusion of the experiment;
- Develop tools for FAIR-by-designing metadata collection and storage. Collaborate with users for the production of adequate metadata for all successfully generated datasets.
- Preserve data for a minimum of 10 years for scrutiny, comparison and reproducibility to reinforce scientific knowledge and integrity;
- Promote data use, after an embargo period, for other scientists in the same field or for cross-disciplinary research and machine learning;
- Enable researchers to mine ELI public data and metadata and in previously unknown ways or apply future methods to existing data
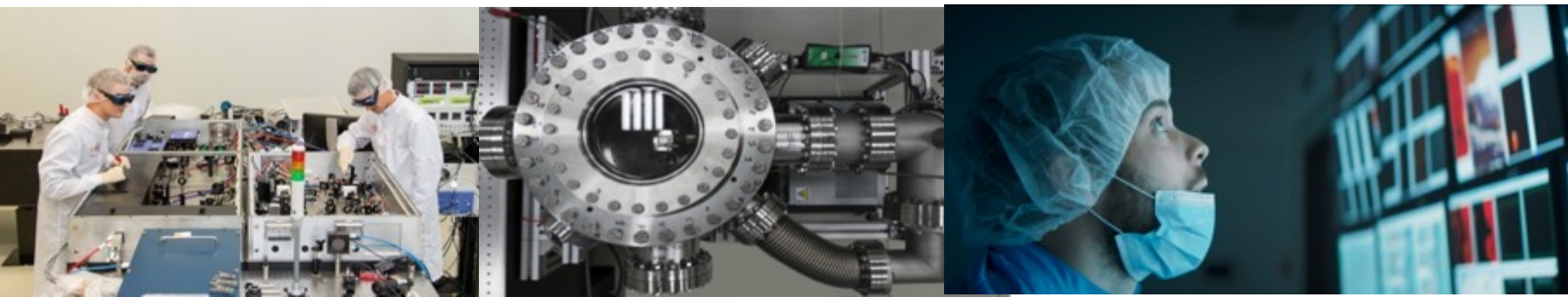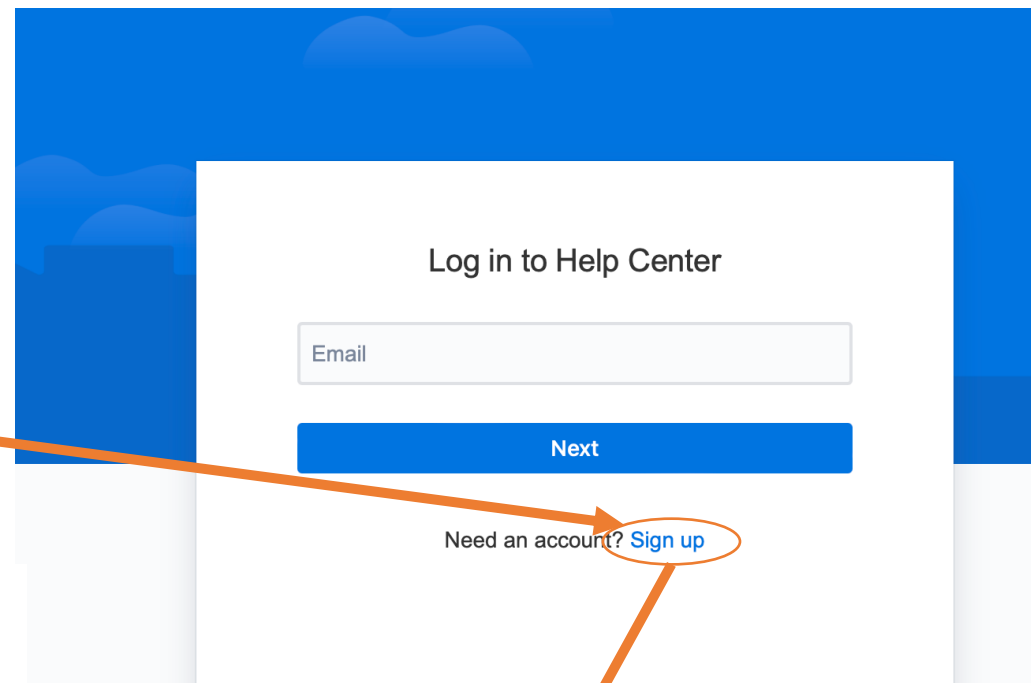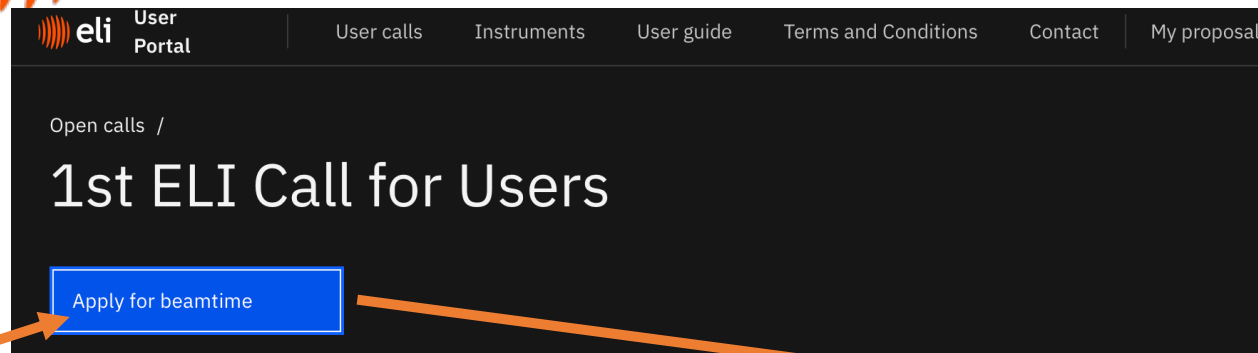
ELI ERIC FAIR Data Policy

- ELI ERIC data shall receive a unique and persistent identifier (Digital Object Identifier). Users shall be able to cite the persistent identifier in any publication that refers to the data.

- A rich metadata format shall be used and associated with the datasets, providing detailed provenance information.

- Data Management Plan (DMP) will be provided by the users applying for experiments.

- ELI ERIC shall be the custodian of the Data, with the responsibility to collect, secure, archive and provide access to the Data.
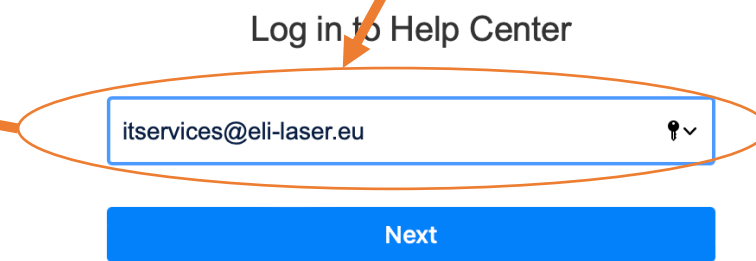
ELI ERIC User Access and PaN community IDp

eli User Portal

User calls   Instruments   User guide   Terms and Conditions   Contact   My proposal

Open calls /

1st ELI Call for Users

Apply for beamtime

Log in to Help Center

Email

Next

Need an account? Sign up

Check your email

We sent an email to itservices@eli-laser.eu. Click the link in the email to finish signing up.

Resend
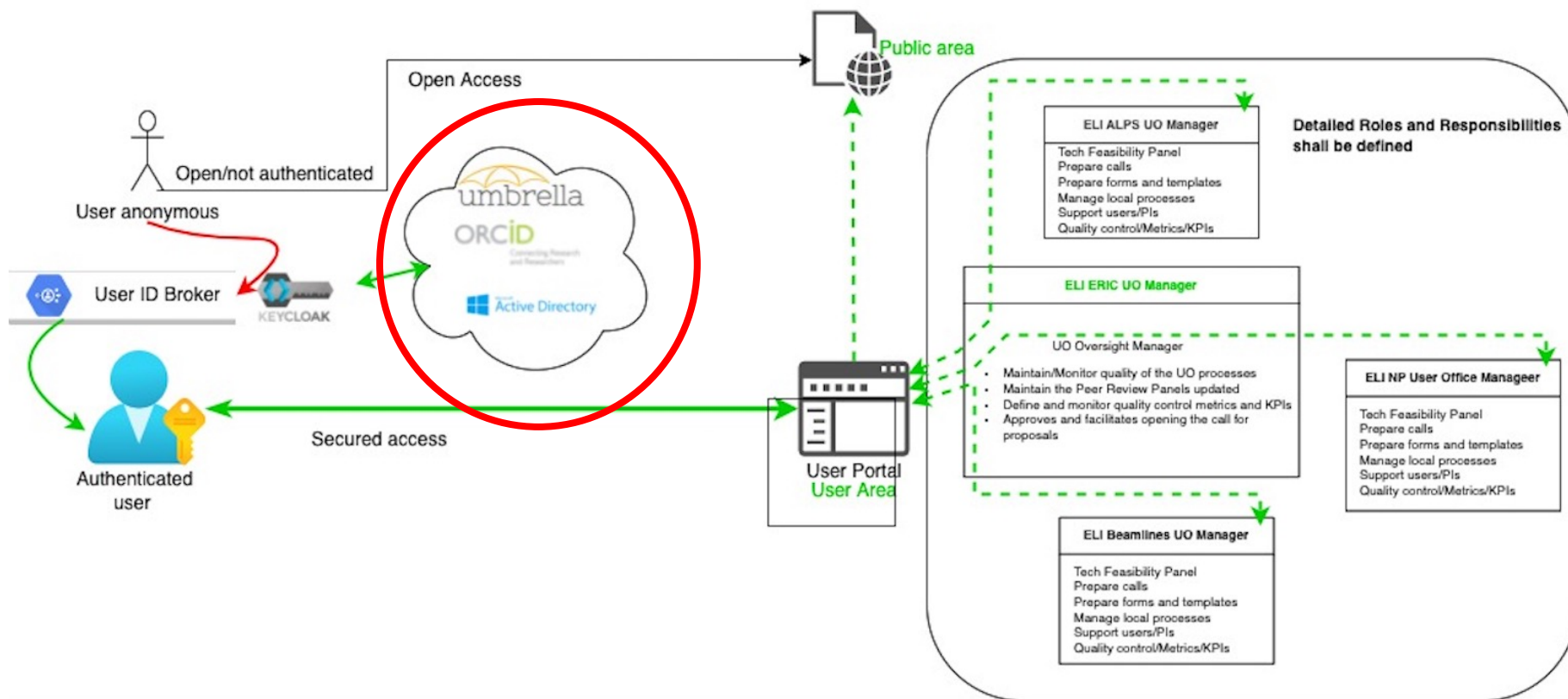
Log in to Help Center

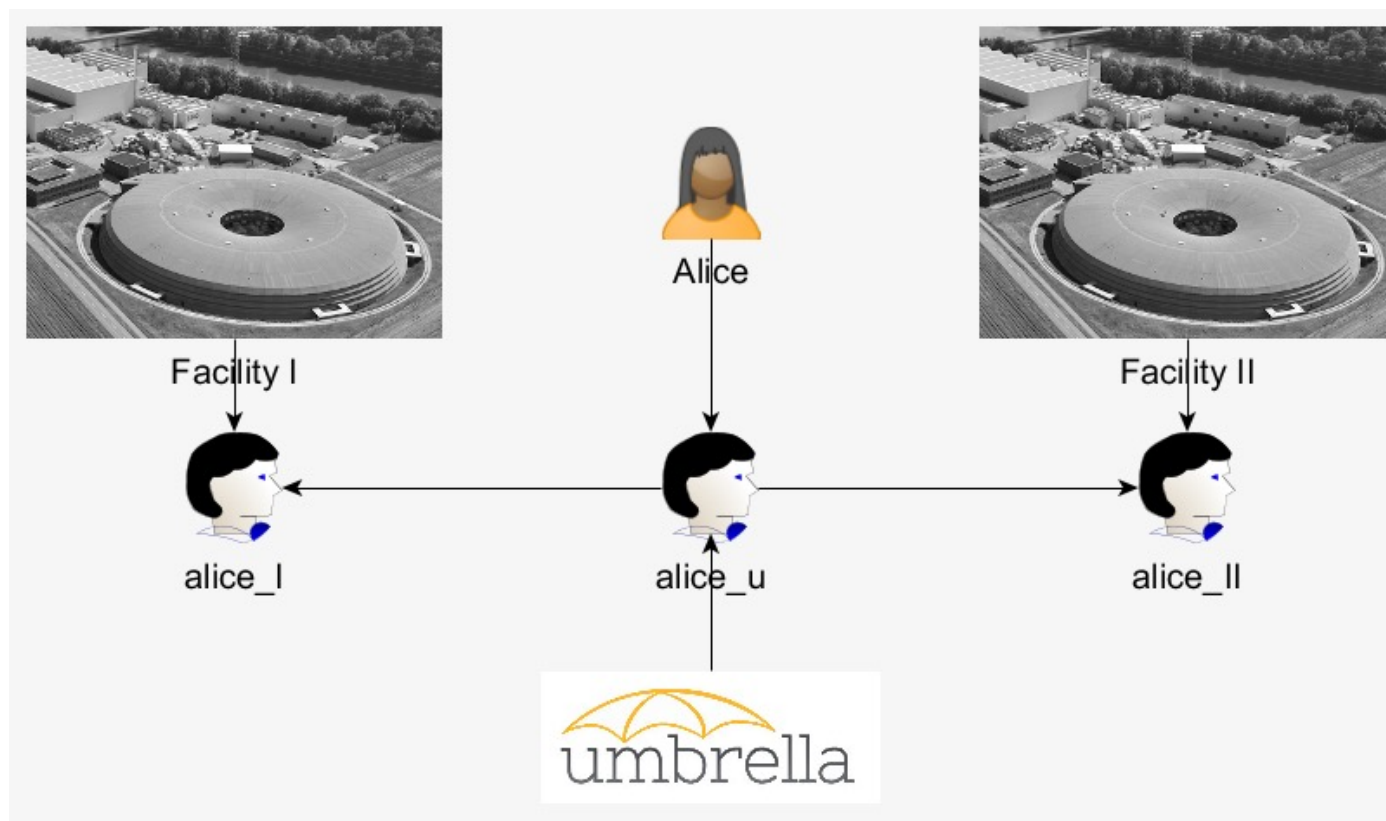itservices@eli-laser.eu

Next

Need an account? Sign up

Already tested and ready for the 2nd Call for Proposals

- Keycloak + Umbrella ID - a Single User ID for all PaN Facilities

ELI ERIC User Access and PaN community IDp

Already tested and ready for the 2nd Call for Proposals
- Keycloak + Umbrella ID - a Single User ID for all PaN Facilities

Already tested and ready for the 2nd Call for Proposals
- Keycloak + Umbrella ID - a Single User ID for all PaN Facilities

First ELI ERIC User Call
ELI ERIC Proposal Submission Form

For a fair and efficient peer-review of your proposal by ELI's peer-review committee, you are requested to fill in the questionnaire below and provide details on the proposed experimental programme and technical requirements of your proposal by using this template. The length of this document, once filled in, should not exceed 5 pages (using Calibri, font size 11).

Raise this request on behalf of *

> TI Teodor Ivănoaica (teodor.ivanoaica@eli-laser.eu)

PI name (First name, Surname) *

> PI1

PI affiliation *

> ELI ERIC

Team members *

> PI Team Member user@facility.eu
> PI Team Member User2@cacility

Instructions: first name, surname, affiliation, citizenship

Filling the application form can be done offline, in case more users should work on the same proposal the PI Should:

- Add them as team members
- Ask each team member to create an account (presented in the previous step)
- Notify the ELI ERIC User Office team or share the proposal with the PI Team members (only after they have created an account)

ELI ERIC User Access

Step 1 – offline

- Download and fill the proposal template
- Add all relevant details and include

Step 1 – online

- Please check all mandatory sections, marked with "*****" and provide the details
- Acknowledge Terms and Conditions
- And Data Processing Rules
- Save the proposal (this stage is saving the proposal)
- The proposal can be still edited

Targeted system for the proposed experiment *****

Scientific and Technical Content of Proposal *****

Drag and drop files, paste screenshots, or browse

Browse

Please upload the Scientific and Technical Content of your proposal using the dedicated template.

Safety requirements: Which of the following does the proposed research involve? *****

- ☐ Animals
- ☐ Biohazards
- ☐ Human subjects
- ☐ Toxic materials
- ☐ Biological samples
- ☐ Explosive materials
- ☐ Radioactive materials
- ☐ None of the above

Terms and conditions and GDPR agreements *****

- ☐ I have read and accept the Terms and Conditions for Access!
- ☐ I have read and accept the GDPR Information Notice!
- ☐ In my capacity as PI, I informed team members participating in this Proposal about the Access Terms and Conditions and GDPR Information, acknowledging that their acceptance of these documents is an admission condition to ELI Experiments!

ELI ERIC Terms and Conditions and **GDPR Personal Data Processing Rules**

Send   Cancel

ELI ERIC User Access
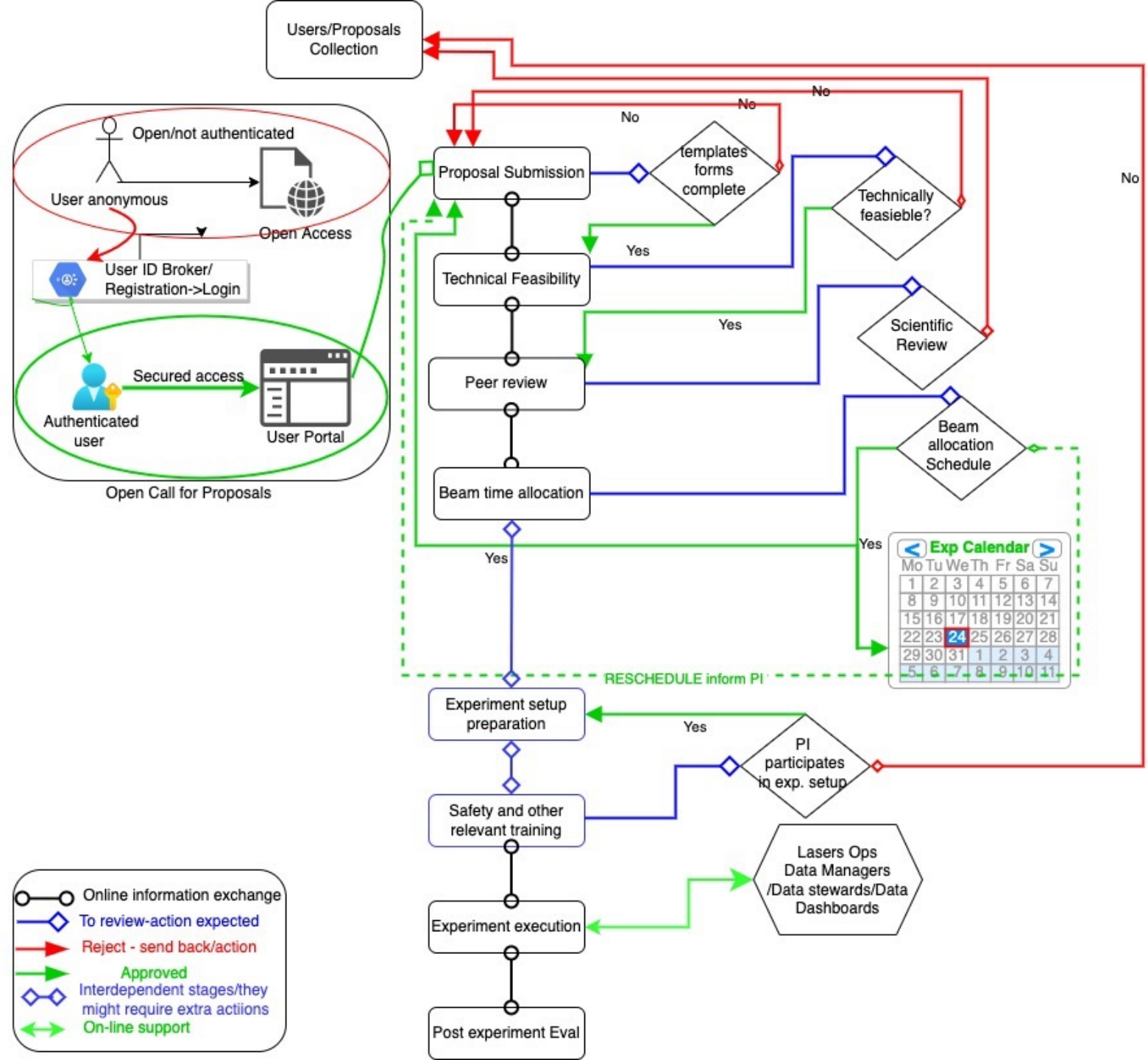
When ready, PI can:

- Access the account and go to "My Proposals)
- Select the proposals and submit for review the proposal to
- Comments are optional, if there are relevant comments they can be added
- After this step, the proposal becomes READ ONLY for the PI Team.

eli

Users/Proposals Collection

Open/not authenticated

User anonymous

Open Access

User ID Broker/ Registration->Login

Authenticated user

Secured access

User Portal

Open Call for Proposals

Proposal Submission

templates forms complete

Technically feasieble?

No

No

No

No

Technical Feasibility

Yes

Peer review

Yes

Scientific Review

Beam time allocation

Beam allocation Schedule

Yes

Yes

Exp Calendar

Mo	Tu	We	Th	Fr	Sa	Su
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4
5	6	7	8	9	10	11

RESCHEDULE inform PI

Experiment setup preparation

Yes

PI participates in exp. setup

Safety and other relevant training

Lasers Ops Data Managers /Data stewards/Data Dashboards

Experiment execution

Post experiment Eval

Online information exchange

To review-action expected

Reject - send back/action

Approved

Interdependent stages/they might require extra actiions

On-line support

Technical Feasibility:

- The objective at this stage is to evaluate the technical feasibility of the proposal within the limits of the campaign specification and facility availability.
- Based on the Proposal Requirements, User Office Manager can add new beam scientists, control systems engineers, IT & computing staff, safety officers etc.
- The panel ensures that the facility can provide the required setup, specified beam parameters, and alignment, including the availability of required computing resources to support the proposed experiment
- This panel advises on the experiment schedule!

Peer Review

- 3 Panellists will perform the Scientific Review of the proposal
- Each proposal receives a score
- Programme Advisory Committee (PAC) meeting will provide the list of accepted proposals

Each approved proposal shall have a DOI!

A DOI, or Digital Object Identifier, is a string of numbers, letters and symbols used to uniquely identify an article or document, and to provide it with a permanent web address

Useful to know:

- How can I use a DOI to find the article it refers to?
  - ➢ If your DOI starts with http:// or https://, simply paste it into your web browser. This will usually lead you to a journal publisher's page for the article.
- You can turn any DOI starting with 10 into a URL by adding http://doi.org/ before the DOI.
  - ➢ For example, 10.3352/jeehp.2013.10.3 becomes https://doi.org/10.3352/jeehp.2013.10.3

Each Data Set provided to a user shall have a DOI!

- Data traceability
- Accountability
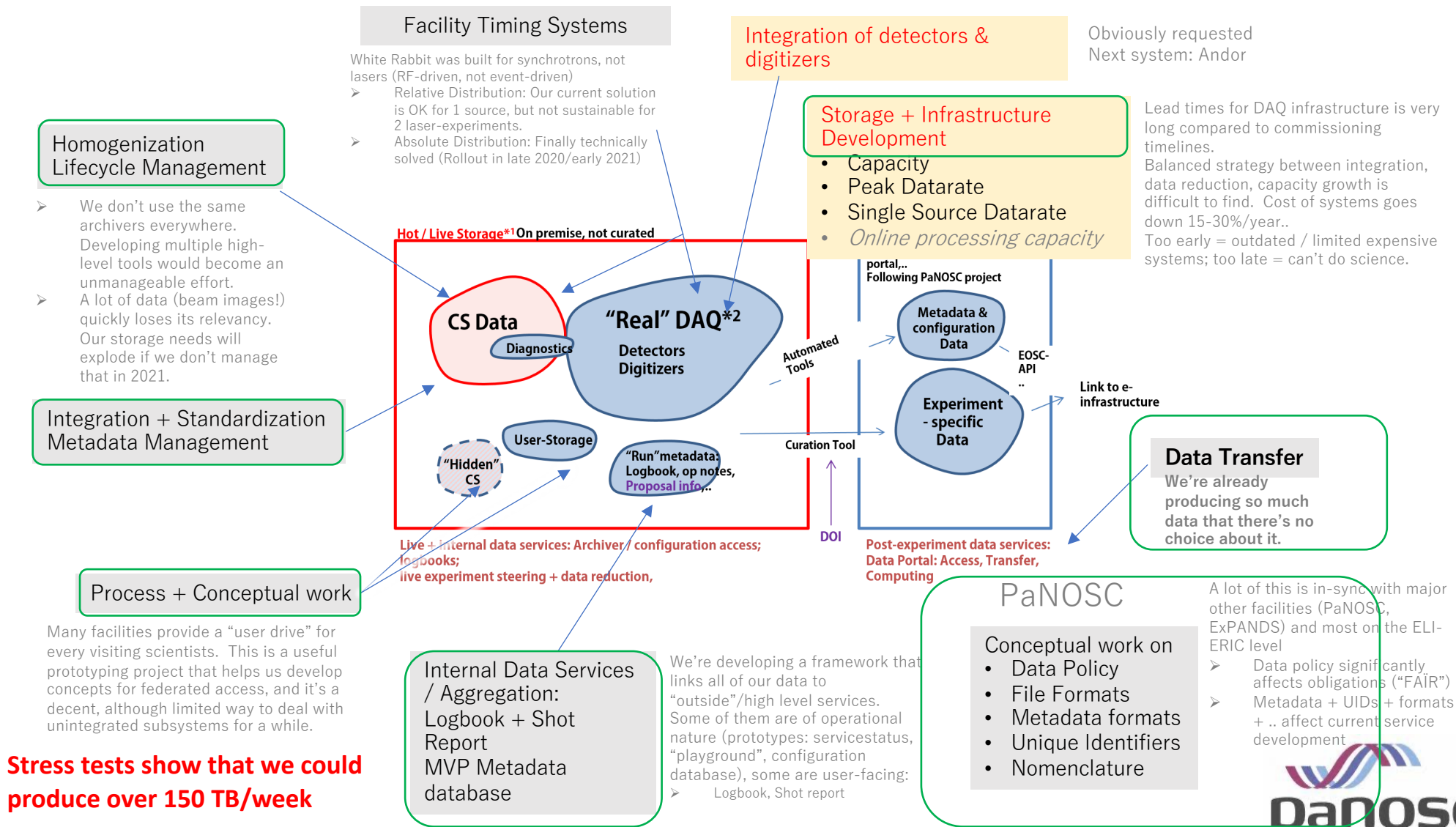- Visibility in the community and outside the community
- ………………

ELI ERIC Data Example

ELI USE CASE!
FAIR is team effort!

Raw = huge disk space
Curated=users' input
Retention depends on raw/curated definition
storage is a considerable cost driver!

Data Analysis tools/methods are needed!

Stress tests show that we could produce over 150 TB/week

Facility Timing Systems

White Rabbit was built for synchrotrons, not lasers (RF-driven, not event-driven)
- Relative Distribution: Our current solution is OK for 1 source, but not sustainable for 2 laser-experiments.
- Absolute Distribution: Finally technically solved (Rollout in late 2020/early 2021)

Homogenization Lifecycle Management

- We don't use the same archivers everywhere. Developing multiple high-level tools would become an unmanageable effort.
- A lot of data (beam images!) quickly loses its relevancy. Our storage needs will explode if we don't manage that in 2021.

Integration + Standardization Metadata Management

Process + Conceptual work

Many facilities provide a "user drive" for every visiting scientists. This is a useful prototyping project that helps us develop concepts for federated access, and it's a decent, although limited way to deal with unintegrated subsystems for a while.

Internal Data Services / Aggregation:
Logbook + Shot Report
MVP Metadata database

We're developing a framework that links all of our data to "outside"/high level services. Some of them are of operational nature (prototypes: servicestatus, "playground", configuration database), some are user-facing:
- Logbook, Shot report

Integration of detectors & digitizers

Obviously requested
Next system: Andor

Storage + Infrastructure Development
- Capacity
- Peak Datarate
- Single Source Datarate
- *Online processing capacity*

Lead times for DAQ infrastructure is very long compared to commissioning timelines.
Balanced strategy between integration, data reduction, capacity growth is difficult to find. Cost of systems goes down 15-30%/year..
Too early = outdated / limited expensive systems; too late = can't do science.

Hot / Live Storage*1 On premise, not curated

CS Data
Diagnostics

"Real" DAQ*2
Detectors
Digitizers

Automated Tools

"Hidden" CS

User-Storage

"Run"metadata:
Logbook, op notes,
Proposal info,..

Curation Tool

portal,..
Following PaNOSC project

Metadata & configuration Data

EOSC-API ..

Experiment - specific Data

Link to e-infrastructure

DOI

Live + internal data services: Archiver / configuration access; logbooks; live experiment steering + data reduction,

**Post-experiment data services:
Data Portal: Access, Transfer, Computing**

Data Transfer
We're already producing so much data that there's no choice about it.

PaNOSC

Conceptual work on
- Data Policy
- File Formats
- Metadata formats
- Unique Identifiers
- Nomenclature

A lot of this is in-sync with major other facilities (PaNOSC, ExPANDS) and most on the ELI-ERIC level
- Data policy significantly affects obligations ("FAIR")
- Metadata + UIDs + formats + .. affect current service development

photon and neutron open science cloud
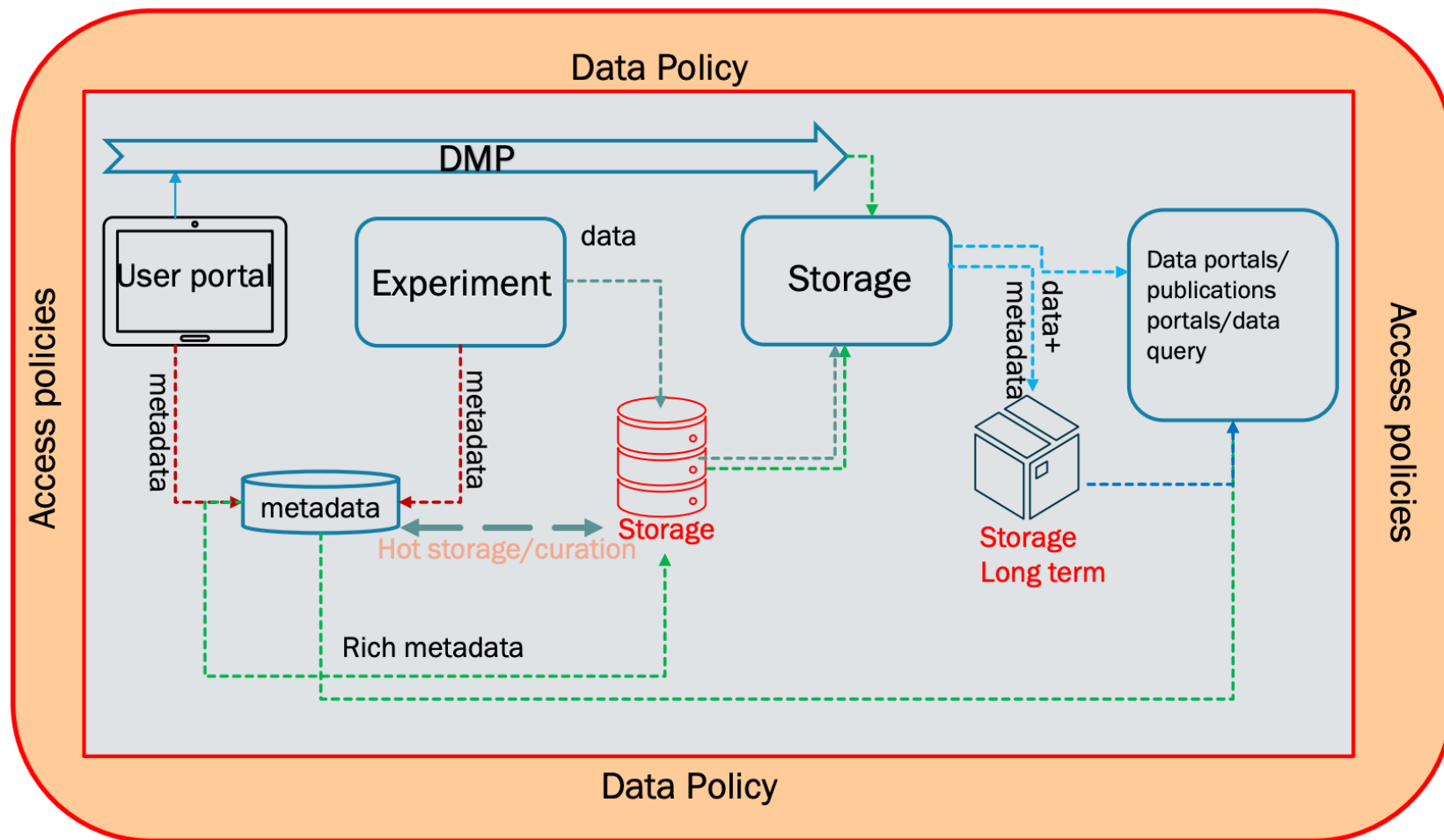
PaNOSC & DMP for users experiments

Definition. The DMP is a 'living' document that supports a researcher through all phases of a project from the planning at the start of the project, through the collection or generation of data and its analysis, to the publication and archiving of data at the end of the project

What it really is: ***The DMP is not (just) a document, the DMP is a Data Quality enabler for the experimental data!***

- Helps users find and understand their data when needed
- Helps avoid unnecessary duplication of data
- Helps document scientific data to ensure ongoing access and continuity when you or other colleagues depart or new research staff start
- Facilitates the validation of the published results
- Ensures your research is visible and has an impact
- Helps a scientist get credit and visibility whenever others cite your work
- Helps share data, leading to collaboration and advances in research
- Helps users achieve compliance with research funder mandates

ELI ERIC DMP, Metadata and Data

DMP is not optional:
- Users can express requirements
- In case requirements are not specified, a default DMP applies
- Default DMP can be updated by users, resulting in a new revision
- Stored with the data

Stage 1

Users' Direct Input:
- Data rates/data types
- Specific details about raw/preprocessed data
- Specific data types and expected volumes

DMP is not optional:
- Users can express requirements
- In case requirements are not specified, a default DMP applies
- Default DMP can be updated by users, resulting in a new revision
- Stored with the data

Stage 2
Users+Facility Team during the experiment setup phase:
- More accurate details
- Full data lifecycle documented
- What is kept/what can be deleted/what intermediary info might be needed
- Data retention policy

User Portal initiates also the DMP

DMP is not optional:
- Users can express requirements
- In case requirements are not specified, a default DMP applies
- Default DMP can be updated by users, resulting in a new revision
- Stored with the data

Stage 3
During the experiment, until the experiment is closed
- Users might require updates, DMP Updates
- Users regularly change the experiments (on the fly) – which will result in a DMP Update
- DMP is a countiuou Data Learning exercise
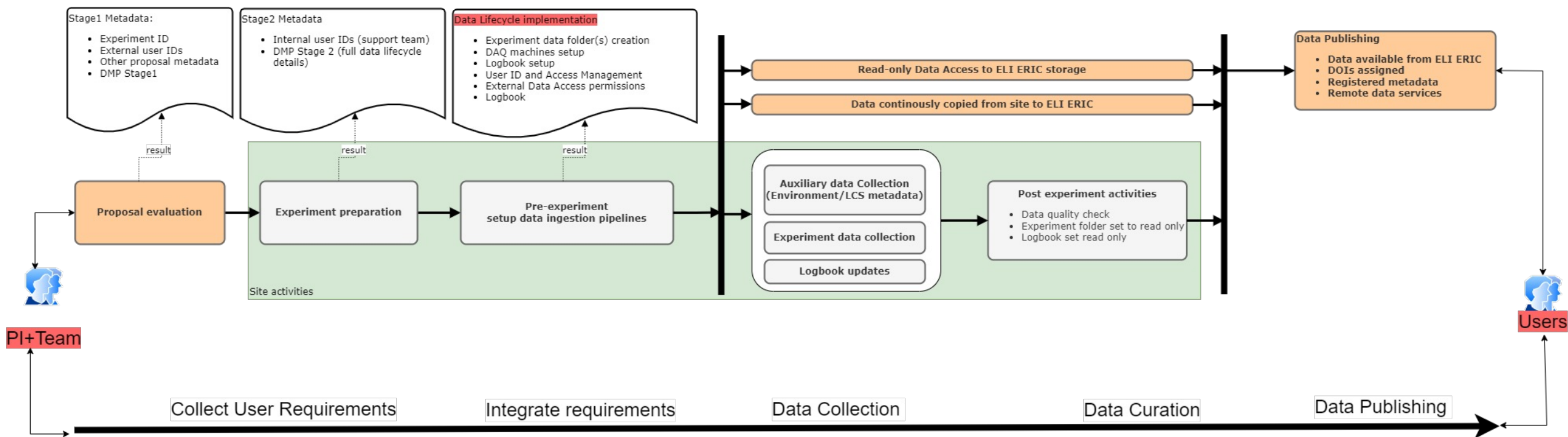
Experiment logbooks, the FAIR Metadata

Data Policy Implementation Better (meta)data, faster science

- Searchable (meta)data and collaborative logbook
- Interactive shot report/logbook
- Remote data transfers and analysis
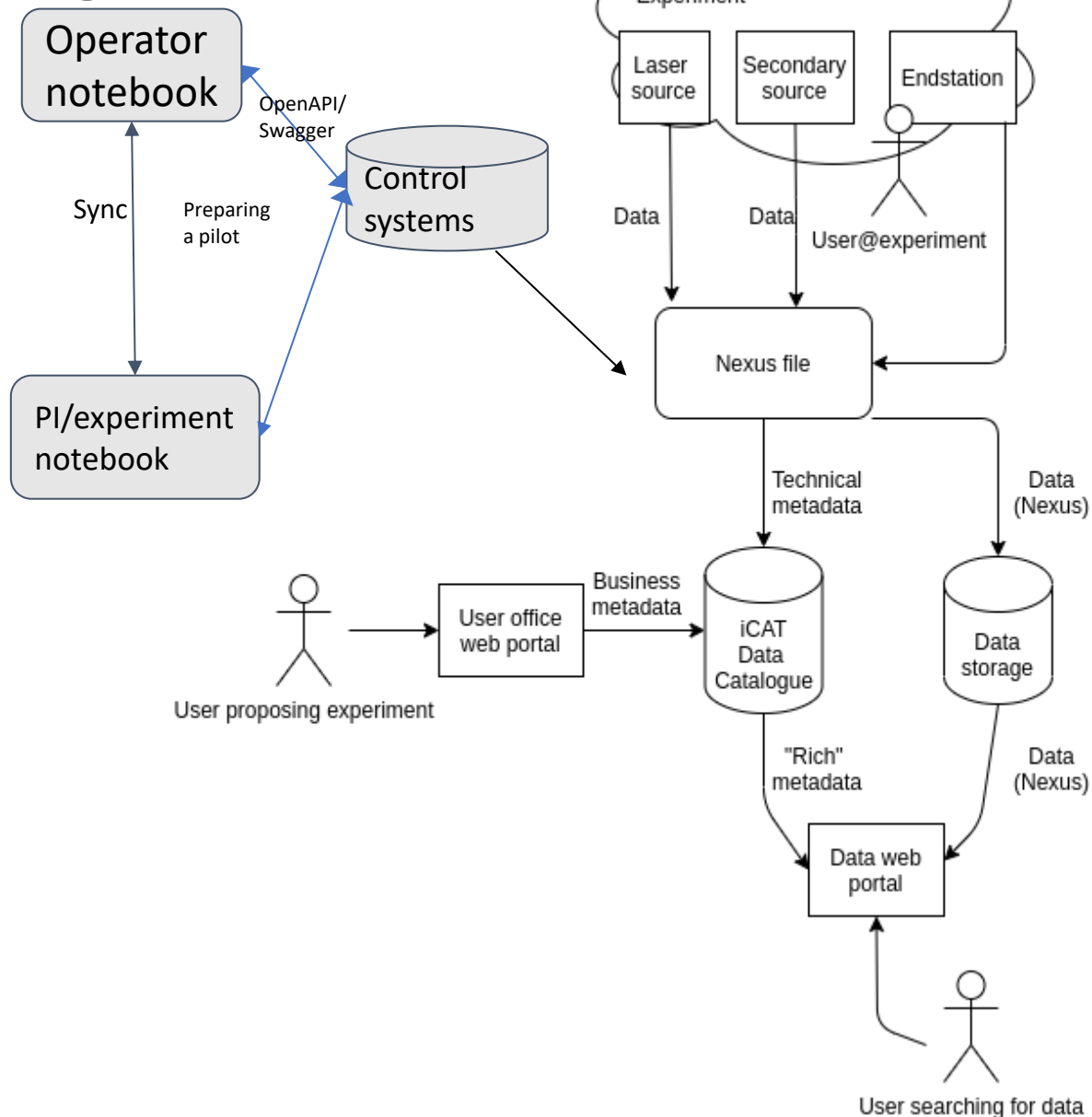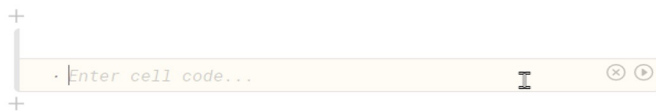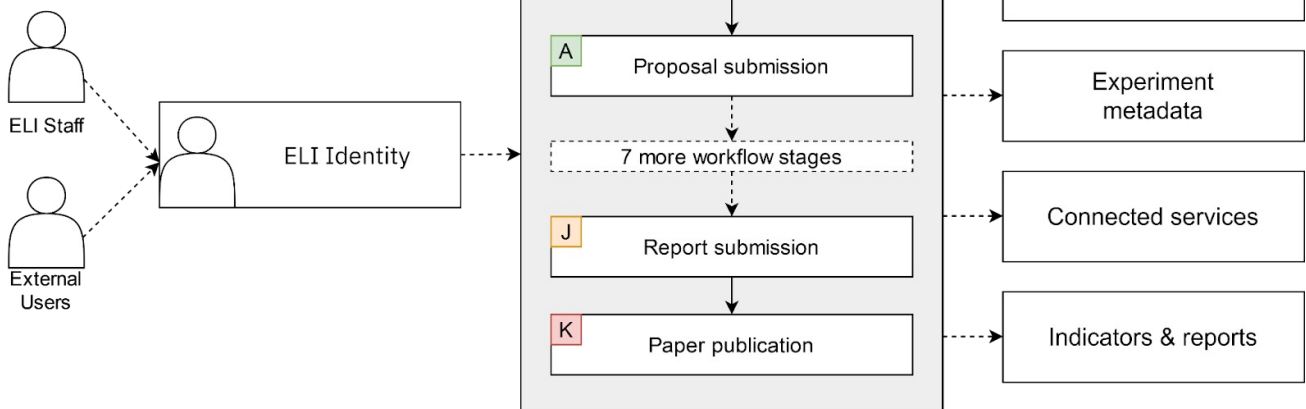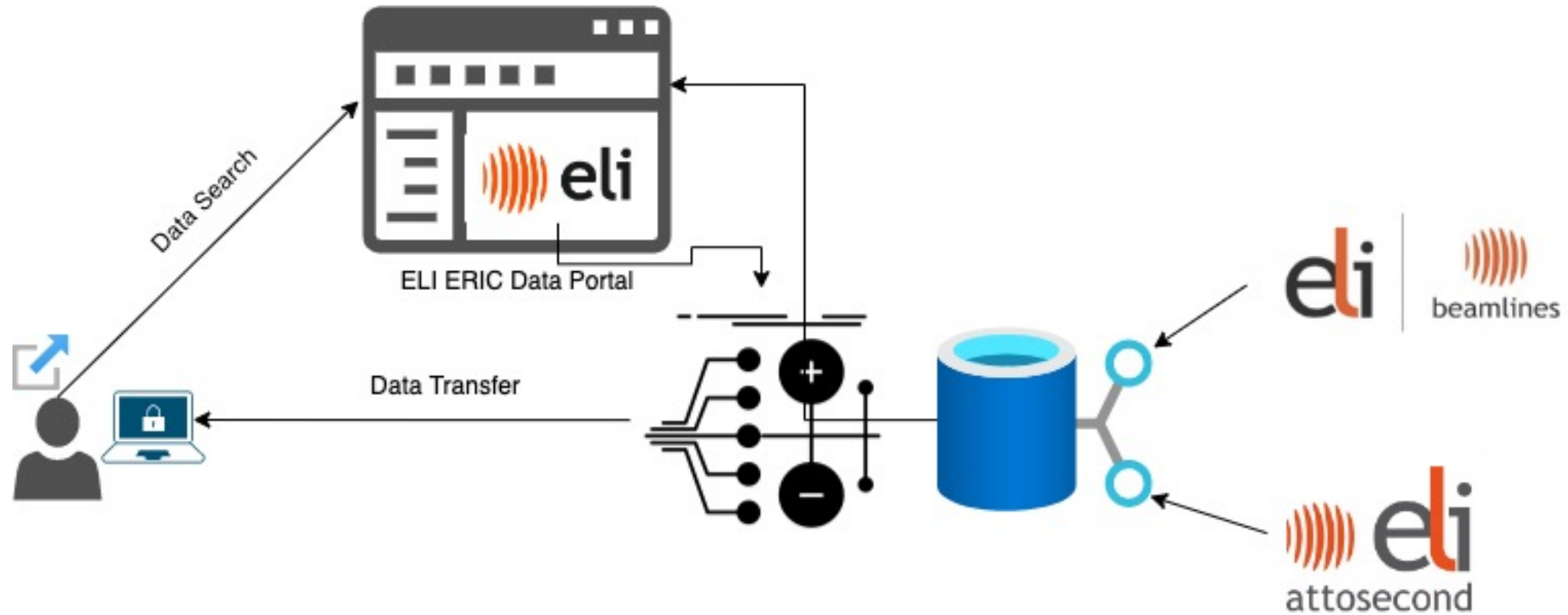
- Data will be made accessible via ELI ERIC Data Portal
- DOIs are assigned per Data set
- Data Transfer from the Facilities to central storage, storing the data under embargo, and after the embargo period
- Data Transfers will be initiated and managed via the Data Portal
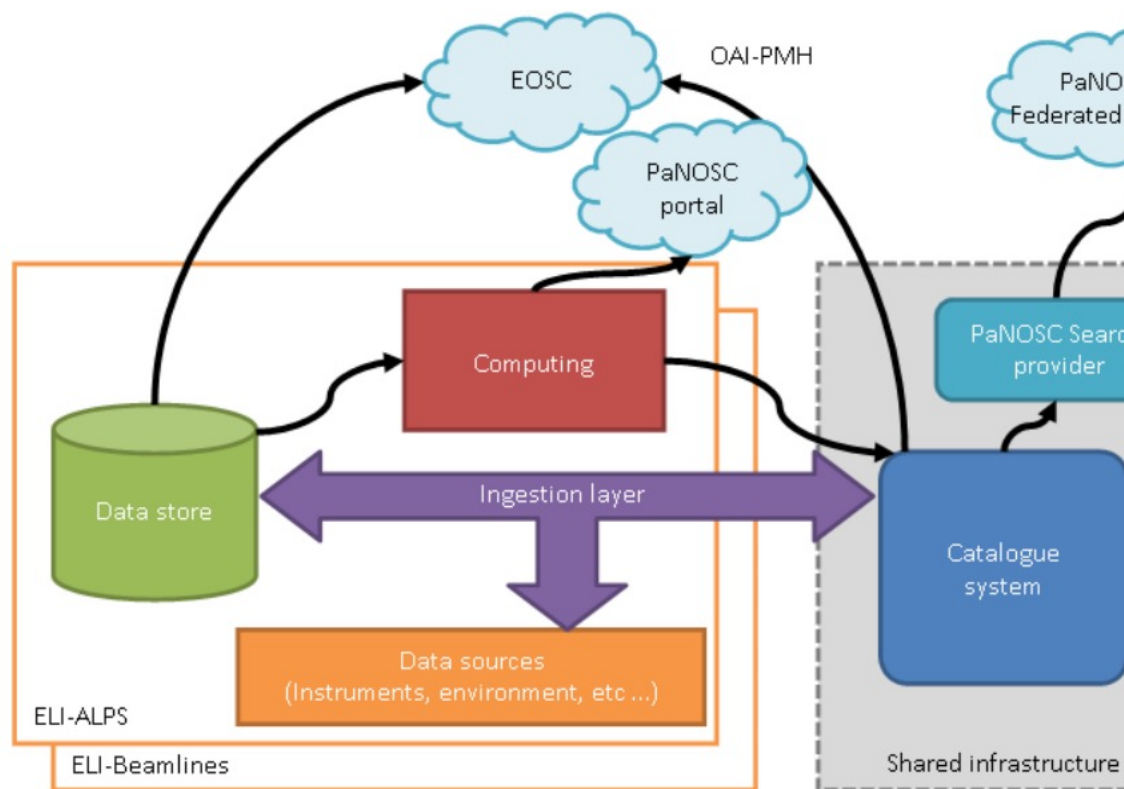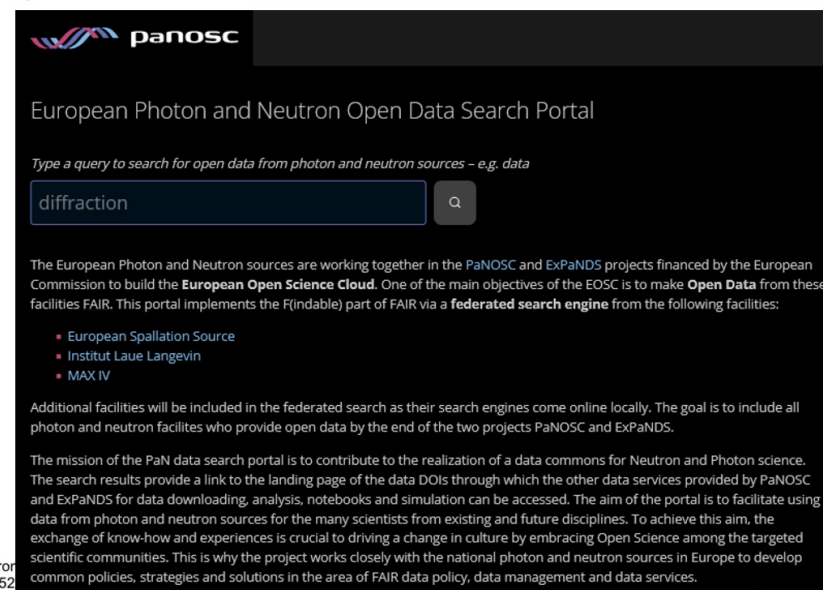
Photon and Neutron Open Data Commons

- Next step is Open Data portals for FAIR Data from PaN sources:
  - Searchable
  - Accessible
  - Downloadable
  - Re-usable

The PaN Open Data Commons will enable new user communities to access and exploit the unique data being produced at the LEAPS facilities to do new science e.g. the Human Organ Atlas is revolutionizing digital histology and medical research with high resolution 3D volumes of complete human organs.

"Data! Data Data! I can't make bricks without clay!"
Arthur Conan Doyle, Writer and Physician