ELI Scientific Data Management System

Teodor Ivănoaica - Birgit Plötzeneder

ELI ERIC – ELI Beamlines

AGENDA

- Data Policy @ELI
- ELI User Journey – The data perspective.
- Data Matters. DMP is the collection of what matters.
- Community and collaborations driving progress.
- Integration challenge – the facility approach.

ELI ERIC Statute

ARTICLE 13 DATA POLICY

13(1) 'Data' refers to all information collected by USERS and the staff while performing scientific experiments under the ACCESS FOR USERS Policy and performing operations of the ELI FACILITIES.

13(2) Open Access to FAIR data sets and metadata stored in Open Access repositories shall be favoured for data collected as a result of the use of the ELI FACILITIES and, to the extent possible in case of software and computer programmes created by the ELI ERIC and the ELI FACILITIES; open source principles shall be considered.

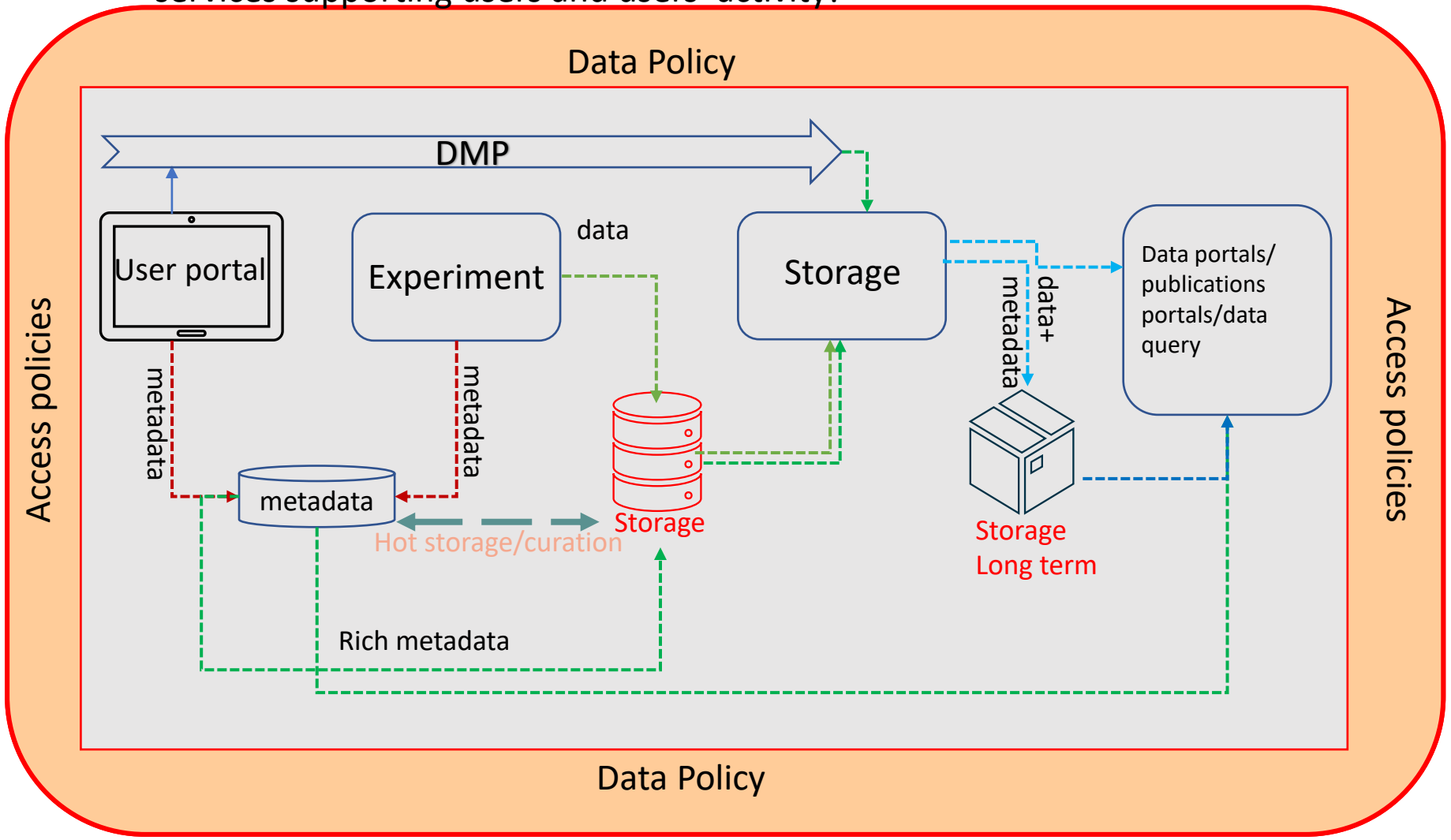
ELI ERIC role, as CUSTODIAN of the Data: *"ELI ERIC shall be the custodian of and steward for the Data, with the responsibility to collect, secure, archive and provide access to the Data. ELI ERIC shall aim at managing Data according to the 'FAIR' principles, meaning that Data shall be Findable, Accessible, Interoperable and organised in Reusable datasets."*

ELI Data Policy *has been developed and will soon be submitted to* the International Scientific and Technical Advisory Committee*. Expected to be adopted by the end of the year.*
"Data Policy governs the management of and access to data relevant to perform and calibrate experiments as well as from experiments performed at the Extreme Light Infrastructure ERIC (ELI ERIC). It pertains to the curation, storage and access to data and metadata collected from the operation and scientific usage of the ELI Facilities."

For a consistent and efficient implementation of the policies, an integrated Scientific Data Management System is needed!

ELI Users' Journey, the Data Perspective!

The Data Policy provides the necessary support to address the above challenges and implement tools and services supporting users and users' activity!



We have the metadata, CS is actively engaging with users to identify/collect/secure:

*__metadata__ is the key for searchable/reproducible data, it does not contain the data but it contains enough information to reproduce a data set/conditions that could be used to reproduce that dataset.(**https://en.wikipedia.org/wiki/Metadata**)

***rich metadata** (https://zenodo.org/record/3862701#.YW-xLxpByUk, https://www.go-fair.org/fair-principles/r1-metadata-richly-described-plurality-accurate-relevant-attributes/)

FAIR (Meta)DATA matters and makes a difference!

Data Policy what it solves what is adding to our work?

What Data Policy solves!!!

1. Addresses data ownership and IPR
2. Address the data-related questions
3. Data lifecycle and data standardizes the way data is treated
4. CONSISTENCY – in our case this is addressed and preserved via the Data Policy
5. Traceability and reproducibility of the data and of thescience

What Data Policy Adds to our Infra:

1. **DMP – collaboration with scientists/users**
2. **FAIR implementation plan**
3. **Standards (NEXUS/HDF)**
4. **DOIs**
5. **Tools Supporting FAIR – catalogues (iCAT/Invenio..)**
6. **Collaboration is the key!**

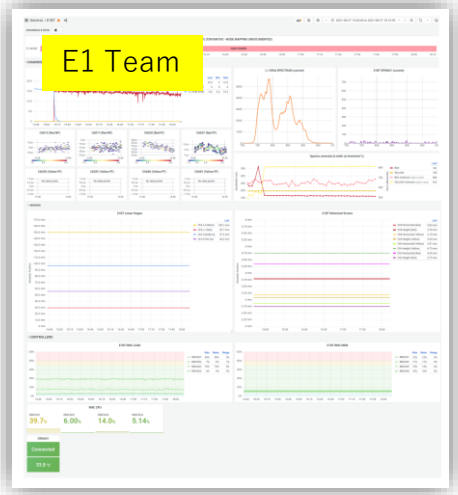
eli beamlines

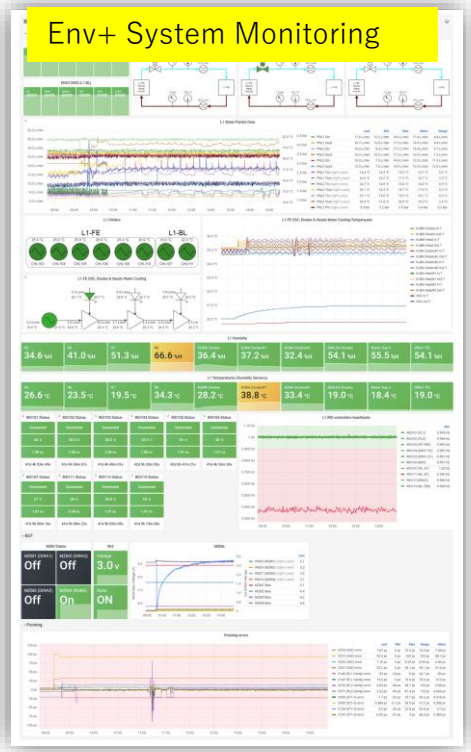
Let's take a look at how much this matters to the scientists.

Access to data matters.
Being able to share data matters.

L1-E1 dashboards
Pre-campaign

1 week later





What is Scientific Data Management? How is this supporting users?

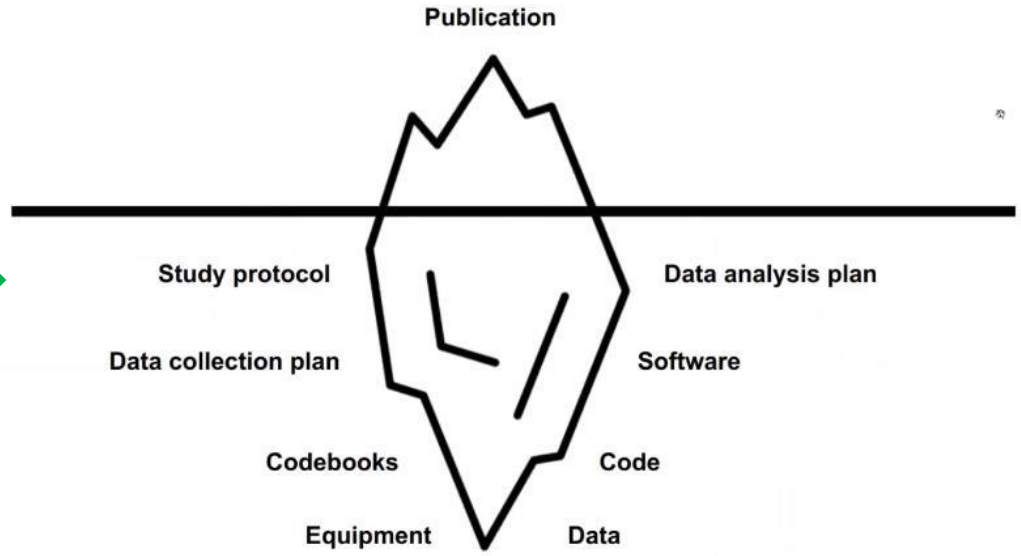
It is ELI's Research Data Management Plan.

It is the **collection of practices** helping us **plan, collect, process, analyse, preserve, share** and make **data re-usable.**

- Data management planning – the **DMP – reflecting the data lifecycle**
- Collecting raw data and metadata – Curation and correlation
- Processing to produce new data
- Analysing data to produce results
- Curating data for the long term
- Sharing data and making it Findable, Accessible, Interoperable and Reusable

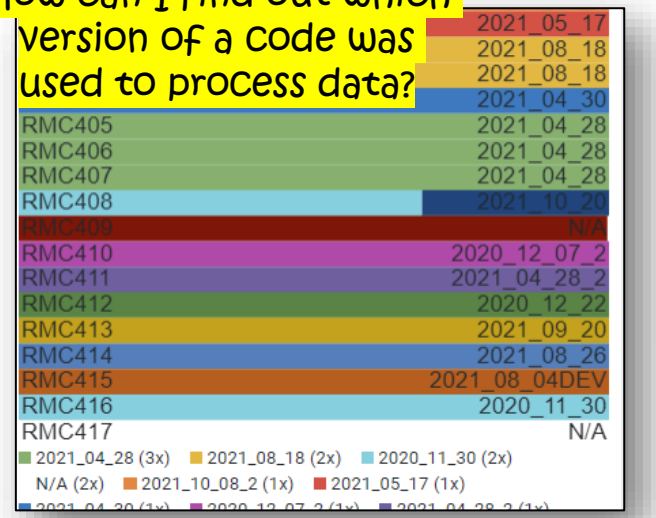
The direct impact of FAIR:

- Facilitates new data
- Derived data produced
- Enables new research
- Accelerates science

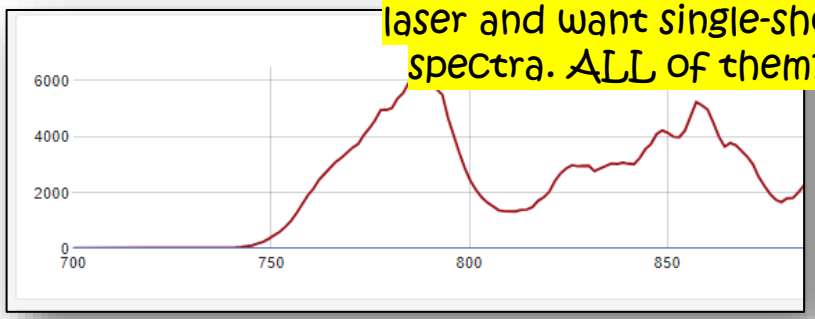


But actually, our DMP is about answering real questions, for example

How can I find out which version of a code was used to process data?

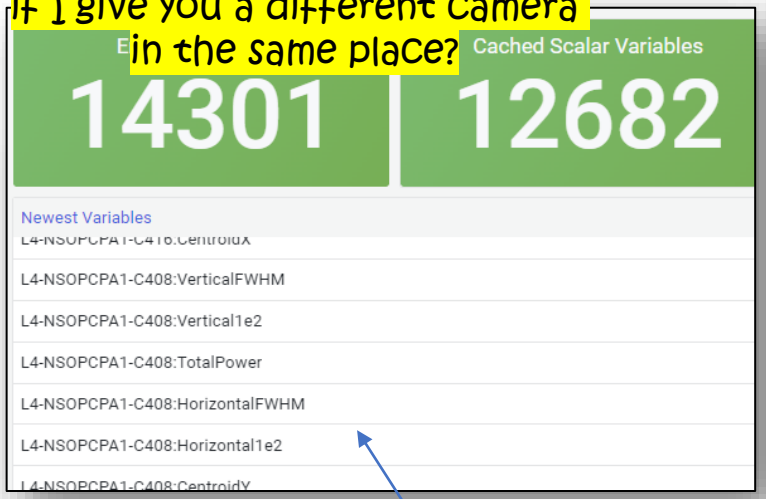


OK, so you have a 1kHz laser and want single-shot spectra. ALL of them?

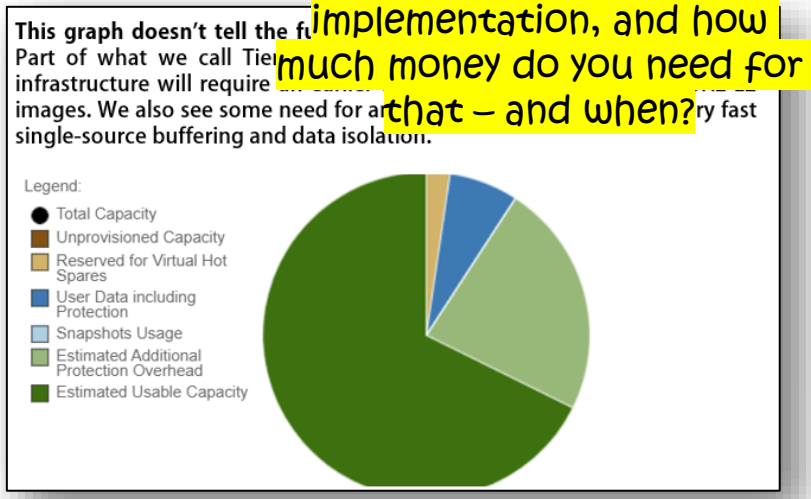


1kHz, 2byte reprate spectral data. It adds up.
It needs, at the very least, language to discuss.

What will happen with my data if I give you a different camera in the same place?



What drives the storage implementation, and how much money do you need for that – and when?

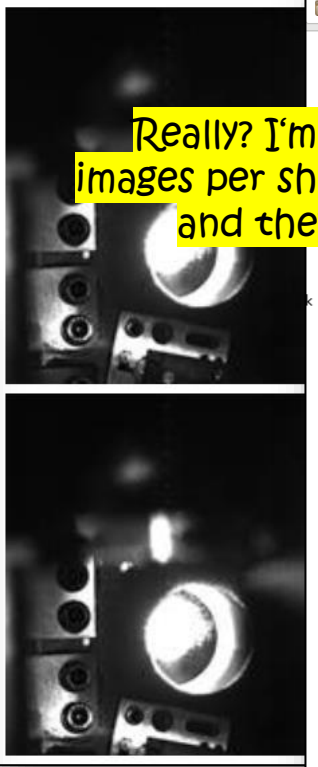


From: Data and data-related infrastructure in 2021, Birgit Ploetzeneder, ELI-BL internal report in 09/2021

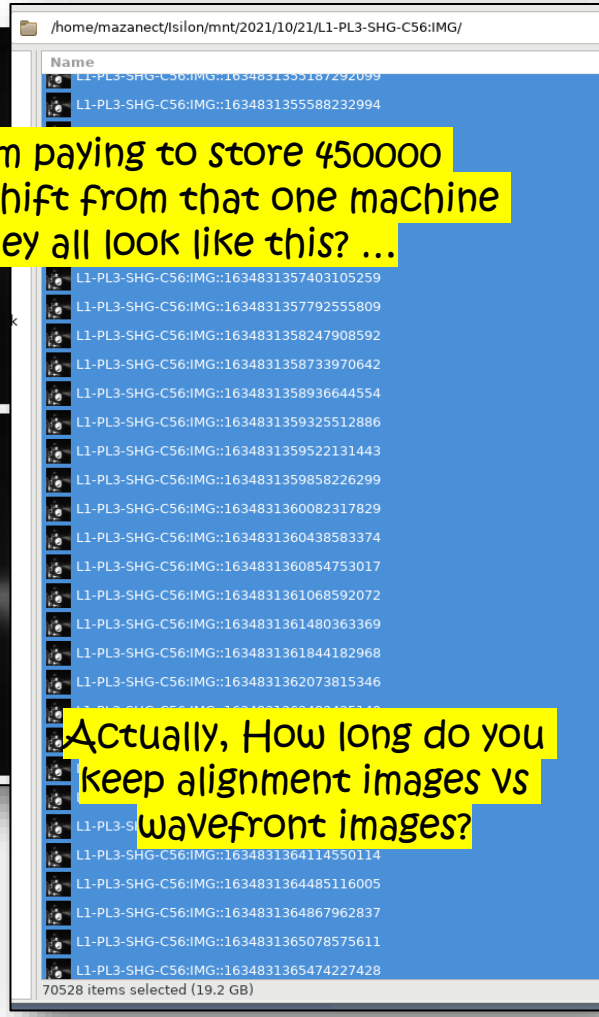
Who knows which power meter is the one, and how do I find out the variable name and what it means?

From: Mazanec Tomáš
Sent: 22 October 2021 14:04:25
To: Tykalewicz Boguslaw
Cc: Majer Karel; Mazůrek Petr; Plötzeneder Birgit
Subject: Re: Killin' L1 RMCs -- continuing webex discussion

As an example: RMC112 successfully served all four of its images
mode. There is 450'000 PNGs archived over 8hrs from it. Boguslav
one = 19GB folder
and some samples ...



Really? I'm paying to store 450000 images per shift from that one machine and they all look like this? ...

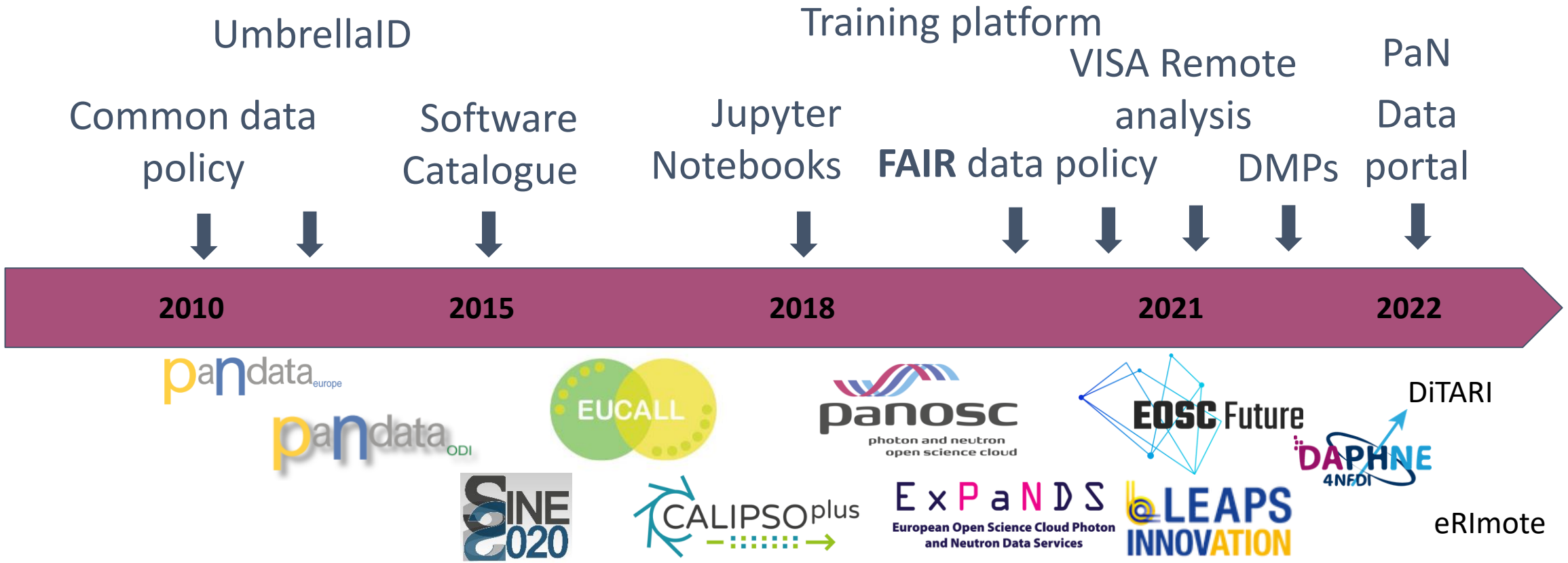


Actually, How long do you keep alignment images vs wavefront images?

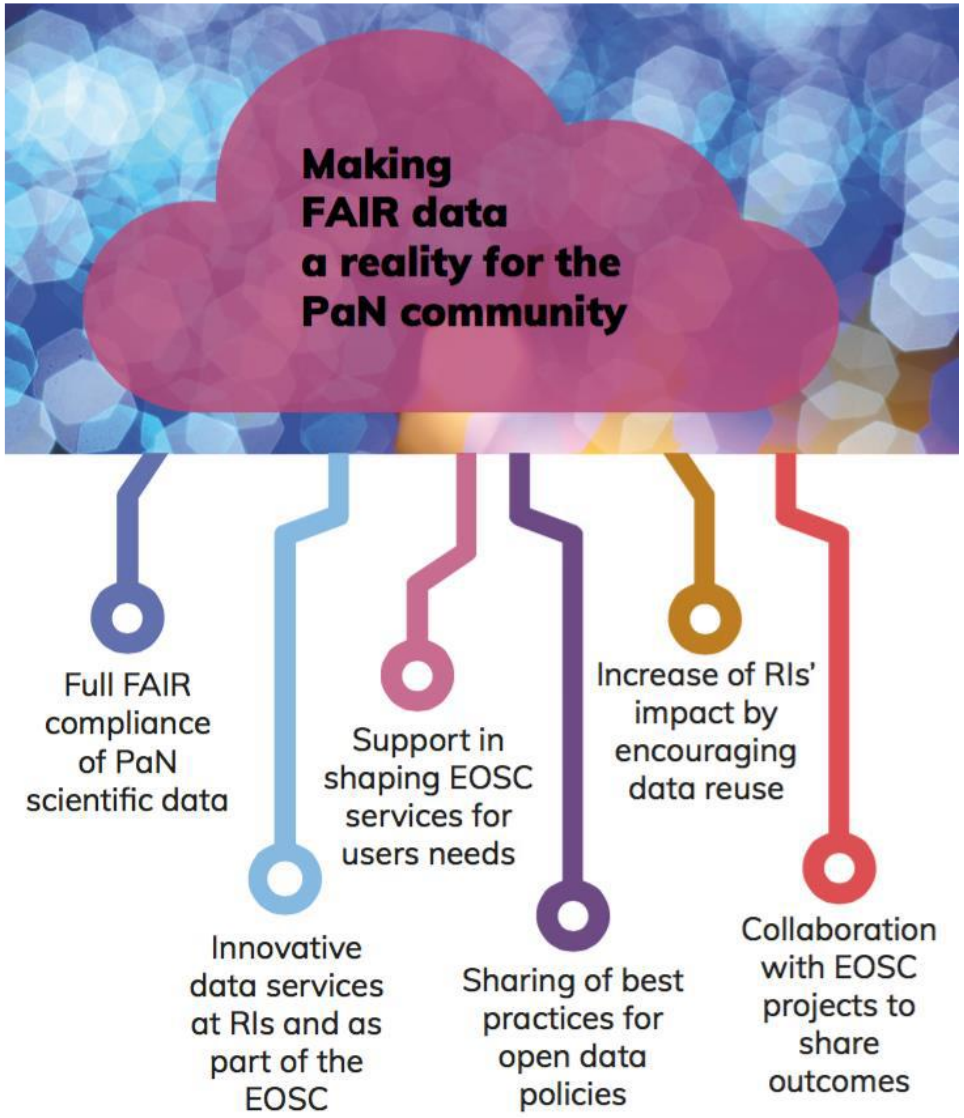
Data stewardship discussion, or: „Just because we made it trivial to archive images now, doesn't mean you can spam my storage! "

Communities support

PaNOSC community we keep building on



PaNOSC is more than tools, is the community sharing the same challenges, same standards and working together to find unique solutions.



What PaNOSC does:

Policies supporting adoption of FAIR policies:

- Data Policy Framework - https://zenodo.org/record/3862701
- Data Policy guidelines - https://zenodo.org/record/4899344

Tools and services:

- AAI
- File Cataloguing solutions and support
- Data tools:
  - Data portal
  - Data transfer tools and solutions for PaN
  - ……

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

Specific ELI Input for implementing FAIR Principles.

IMPULSE Project Goal: A global platform for high-power laser science and development, uniting the facilities of the Extreme Light Infrastructure together.

ELI-ALPS
ELI-HU Non-Profit Ltd.

ELI-NP / IFIN-HH
Institutul National de Cercetare-Dezvoltare pentru Fizica si Inginerie Nucleara – Horia Hulubei

FORTH
Idryma Technologias Kai Erevnas

FZU
Institute of Physics of the Czech Academy of Sciences

HZDR
Heimholtz-Zentrum Drezden-Rossendorf EV

INFN – LNS
National Institute of Nuclear Physics – Laboratori Nazionali del Sud

IST
Instituto Superior Tecnico

LMU
Centre for Advanced Laser Applications

QUB
The Queen's University of Belfast

STFC
Science and Technology Facilities Council

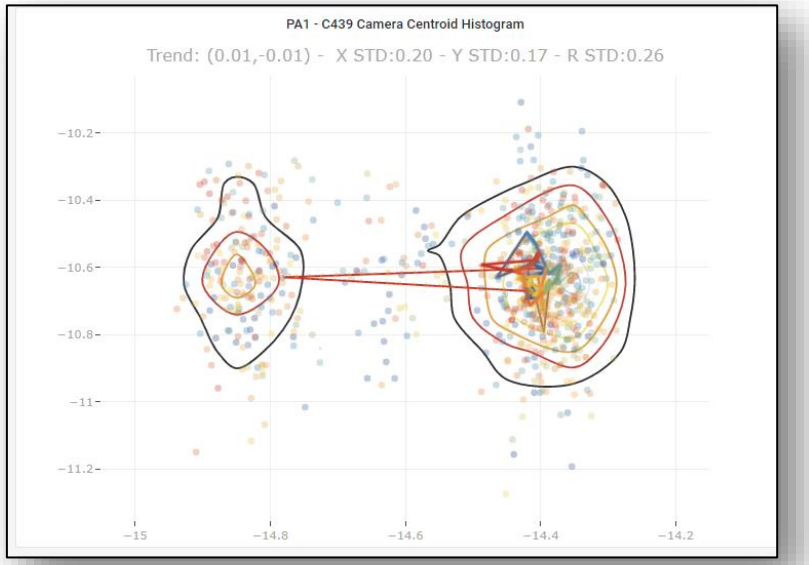
TUD
Technische Universität Darmstadt

Provides the necessary support for having the FAIR principles and all tools and services implemented based on ELI Specific requirements.
Major outcomes that are already used in the design:

- Users office workflow and user portal processes - supporting the implementation of the **DMP;**
- **Simulation software expected to improve operations – supporting the data analysis and simulation services for users**;
- CS teams are joining efforts – accelerating the development of data tagging, data correlation and data curation processes;
- …….
- Most of the activities are boosting the design and implementation of the Data Policies and data services.

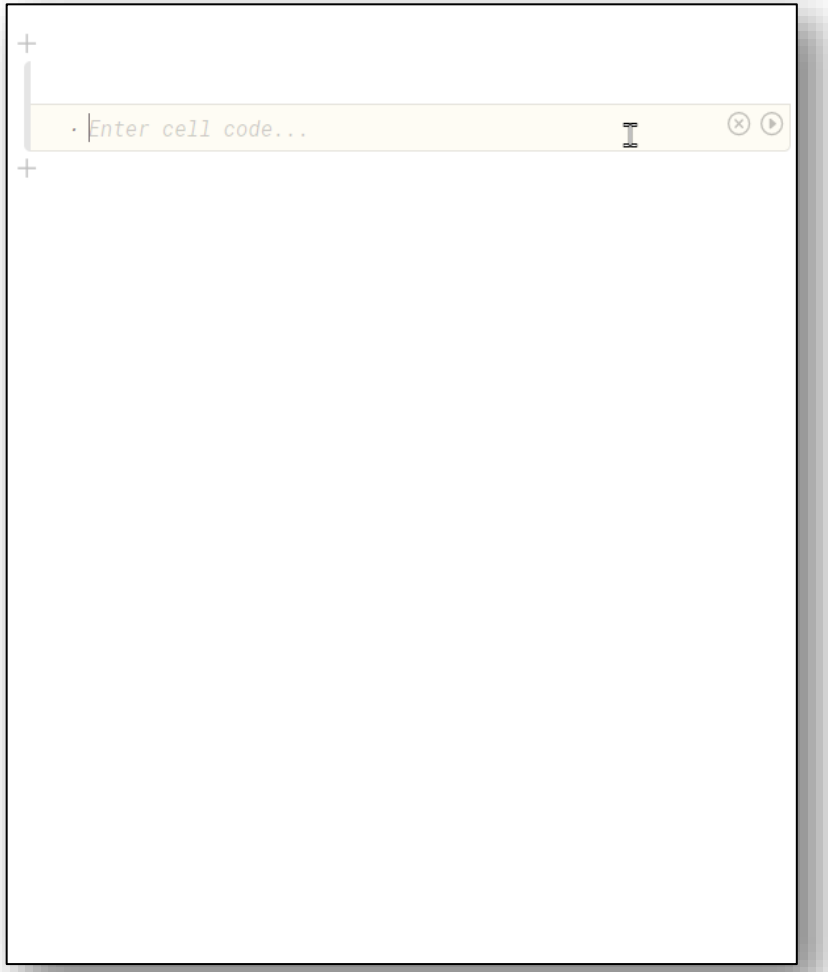
ELI ERIC Electronic Logbook project, started with one PhD student is ready to be tested:
A pilot is discussed with ELI Beamlines CS, another pilot to be started also @ELI ALPS
Inspired by PaNOSC WP 3 Meeting

"Google Docs with scripting , math and archiver integration"



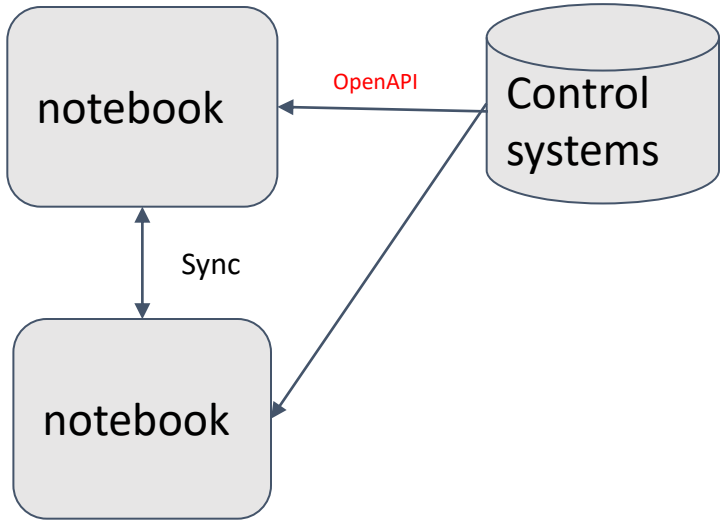
OpenAPI is supported by ELI-BL CS standards by default, and arbitrary interfaces can be generated on both sides. This will allow very creative utilization and not constrain it to access of historic data. While on-prem, a scientist will be able to create live scripts that might even actively control parts of the CS, outside the lab that can be tested with simulated data, and *the same code and visualization* is then part of the logbook.

There's a lot of work to be done architecturally – API and access scope definition are key tasks!

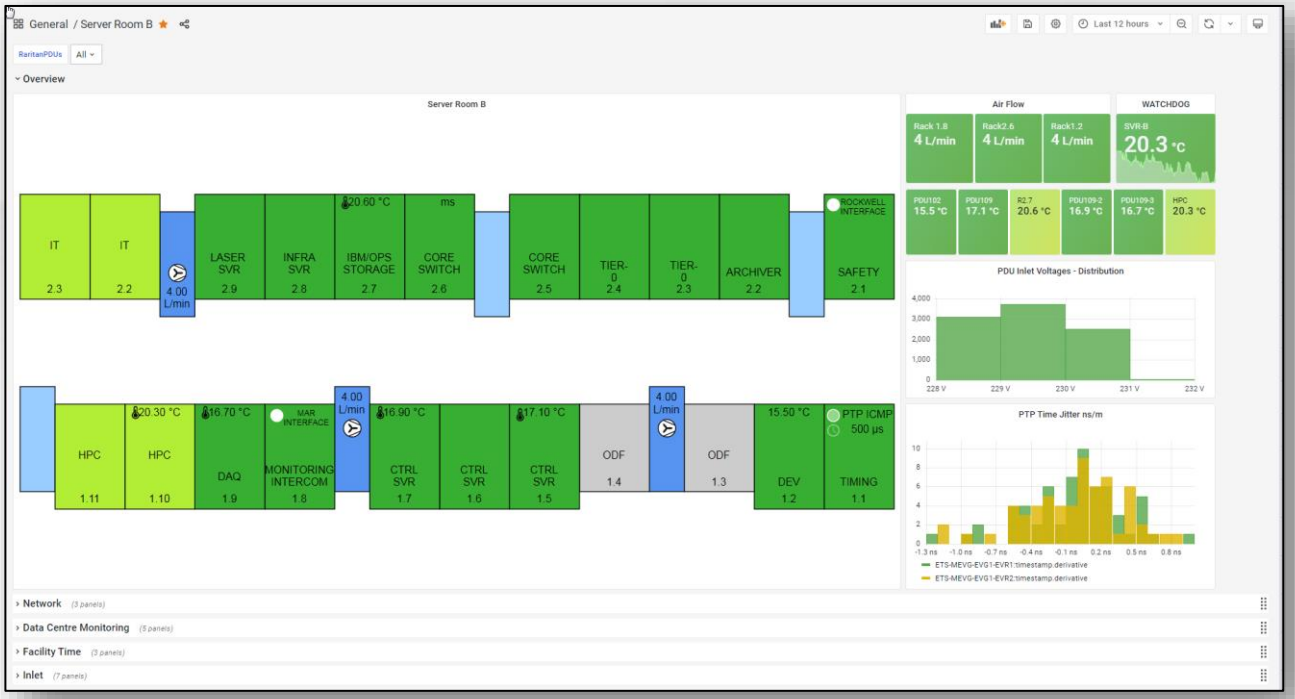


Pilot addressing requirements:

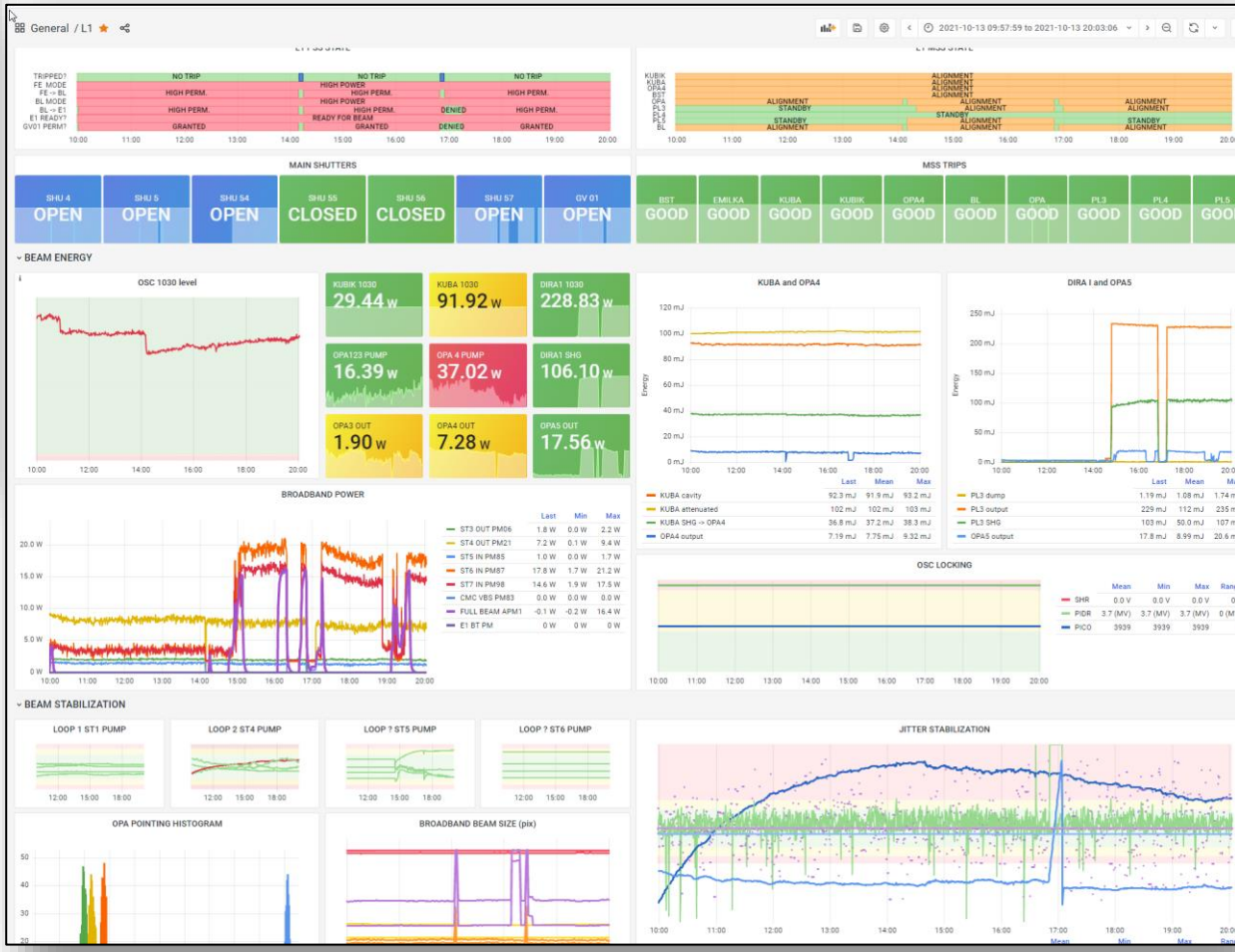
- Distributed, web-based real-time collaboration platform
- Notebook format (scripting!)
- Live Data access



LPOM = yes, we want to add ML/AI to that..



Live Demo.



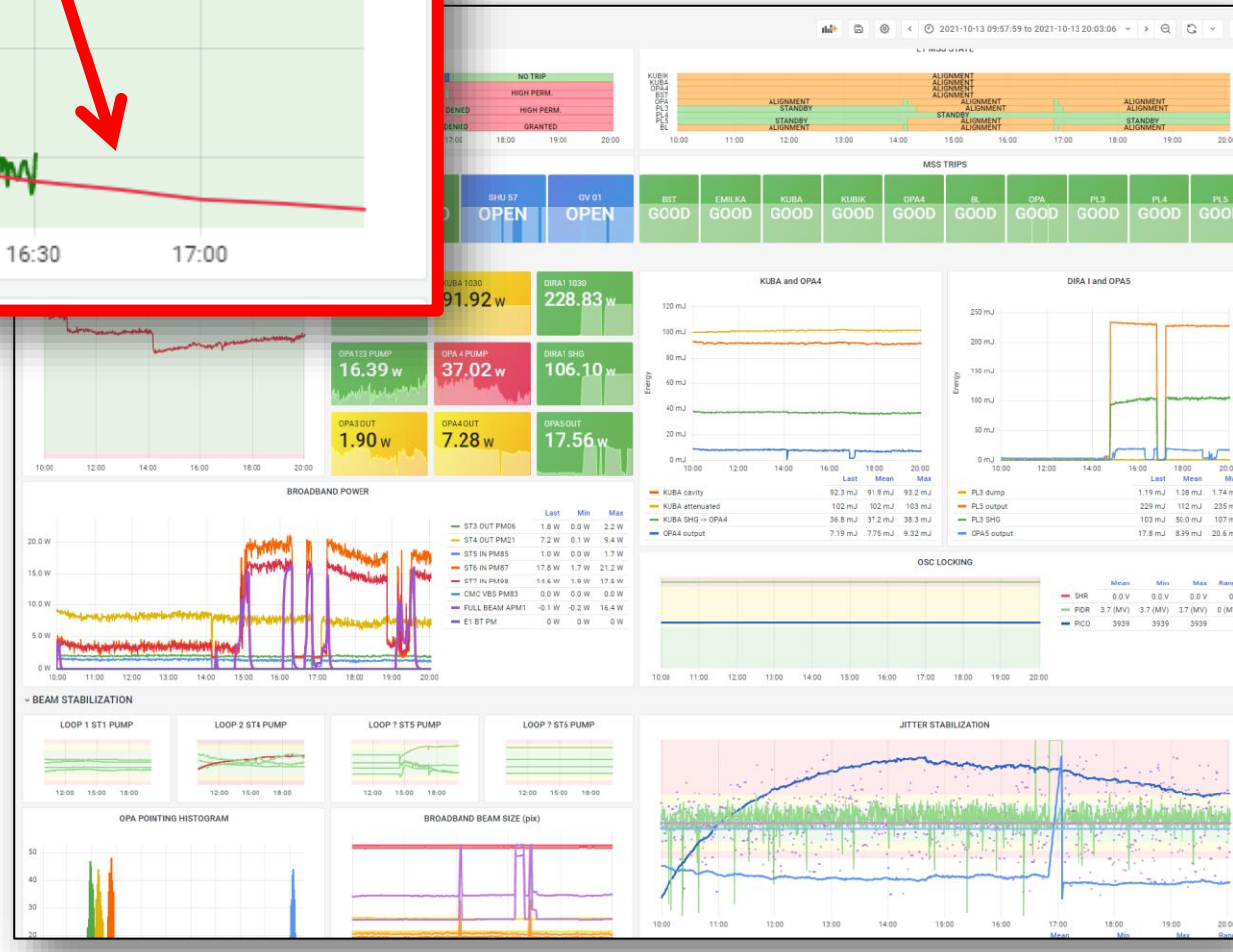
Combination of all data in one place:

- Monitoring (Zabbix..)
- Measurements (Archivers..)
- Aux Systems (Building,..)
- Behavioral data (Logfiles)

Feasibility study & Prototype for Elastic (+Grafana)

LPOM = yes, we want to add ML/AI to that..

(and add the human into this loop via the logbook.)



Combination of all data in one place:

- Monitoring (Zabbix..)
- Measurements (Archivers..)
- Aux Systems (Building,..)
- Behavioral data (Logfiles)

Feasibility study & Prototype for Elastic (+Grafana)

From: Plötzeneder Birgit <Birgit.Plotzeneder@eli-beams.eu>
Sent: Wednesday, October 20, 2021 2:03:21 PM
To: Naylon Jack Alexander <JackAlexander.Naylon@eli-beams.eu>
Subject:

How to fry a digitizer.. "The fan was annoying"

Scientists..

Main current data challenges

There‘s an element of truth in this joke: We‘re not doing this for ourselves.

Support Frameworks for User DAQ

- Campaign periods *can* be very short (1-2 weeks)
- *Sometimes* high flexibility of setups („a station has 30 devices, 10 of them are used in a campaign, and the users bring 3 random new ones and once a year the group finds money for a new detector“).
- Beam and configurations, and *relevant* parameters can be highly dynamic and dependent on use-case! Shot plans adapted day by day..

Limited support time, finite resources: **How do you support this kind of DAQ best and turn data FAIR without constraining science?**

Projects: Processes and tools!

- „Rapid / hybrid“ integration – scripted data insertion of arbitrary data into the archivers = data pipeline + powerful logbooks
- „We need something as simple as Dropbox to upload“
- Provision of typical solutions anticipating support needs

Good: This can be used and developed for internal stakeholders with lower integration degree
Bad: „My IT guy at home prevents me from setting my laptop to facility time and I can‘t install this software on yours.“

Main current data challenges

Let‘s talk about metadata!

Or: What should my shot report look like?

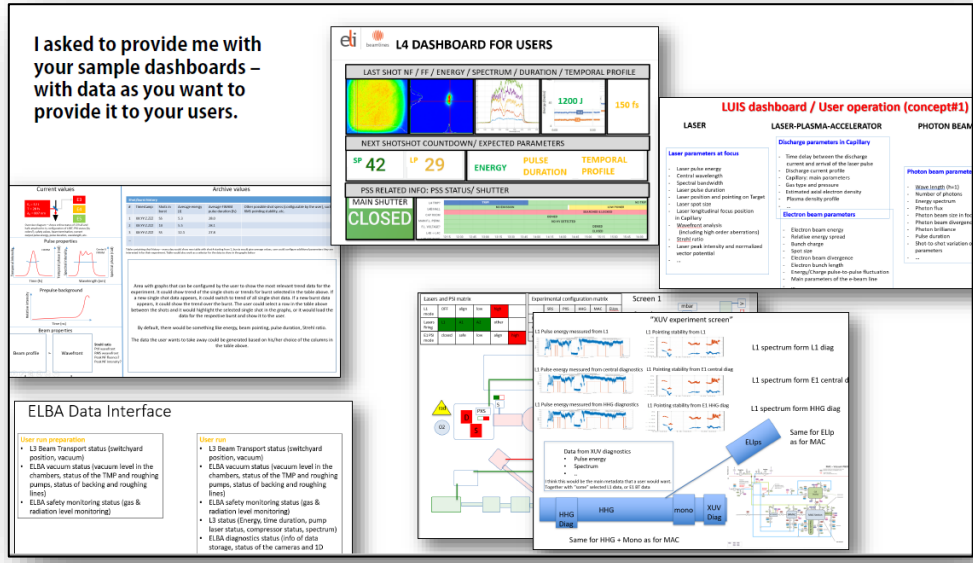
- High diversity and sometimes operational flexibility at the end-station
  - Serving different communities - support all standards?
  - Integration challenges based on operational modes
- More potential – perhaps less value - on target + laser diagnostics side: Many useful projects to define targets and especially primary source diagnostics and create metrology standards
- Some common patterns in many places: Shot reports, runs / bursts,..

Questions to the community

- Standards for common parameters especially on laser side:
  But also vocabulary for principles: In-situ / on a leak beam, single-shot /rep rate /accumulative,..
- Are there commonly used examples for shot report structure?
  (Especially interested in: Shot numbers and tags for synched beams / pump-probe; tagging based on operational modes)

Material	Type	Purpose	Applications	Specifications	
Solid	Thick	Plasma Mirror (PM) Interaction Secondary Source (Seso)	Higher Harmonic Generation EMP	Material Operational Mode	AR coated glass (silica) 1 kHz, 10 Hz Single Shot Size 20 CM dia circular
	Foil	SeSo Interaction	THz (Generation, Samples) Particle Acceleration X-ray, XUV EMP	Material	Al, Au, Ta, Ti, PE, Deuterated PE Diamond like Carbon Nano and micro structured Targets Aerosol Targets (Foam)
				Thickness	50 nm -> 12 um
				Target Holder	Matrix Elements

Result of an IMPULSE Target Survey



ELI-BL Internal Data Workshop: „What data do you want to provide to your users, and in what format?“ „How should a shot report be structured so it would be easy for you to process the data?“

THANK YOU

ELI ERIC Computing Team
Teodor Ivănoaica

ELI Beamlines CS Team
Birgit Plötzeneder