ELI FAIR Data challenges Federated services for Open Science

Outline

ELI as DATA Facility

ELI FAIR Data Services

- **Federated services**
- **Open Data**
- **Open Science**

ELI as a DATA Facility

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)

Integration of detectors & digitizers

Obviously requested
Next system: Andor

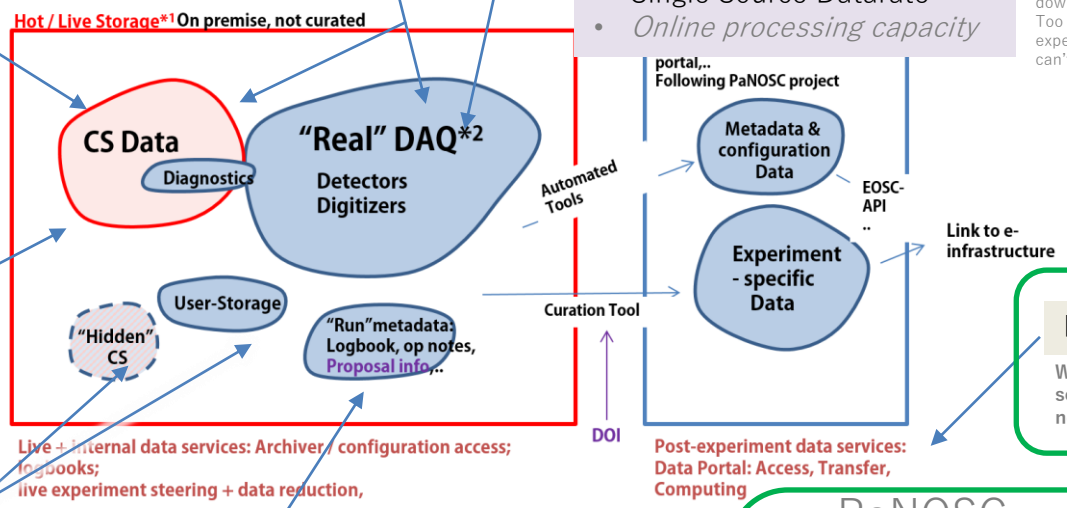
Storage + Infrastructure Development

- 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Too early = outdated / limited expensive systems; too late = can't do science.

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. If we don't manage that in 2021, our storage needs will explode.



Integration + Standardization Metadata Management

MVP Data Transfer

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

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.

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

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

PaNOSC

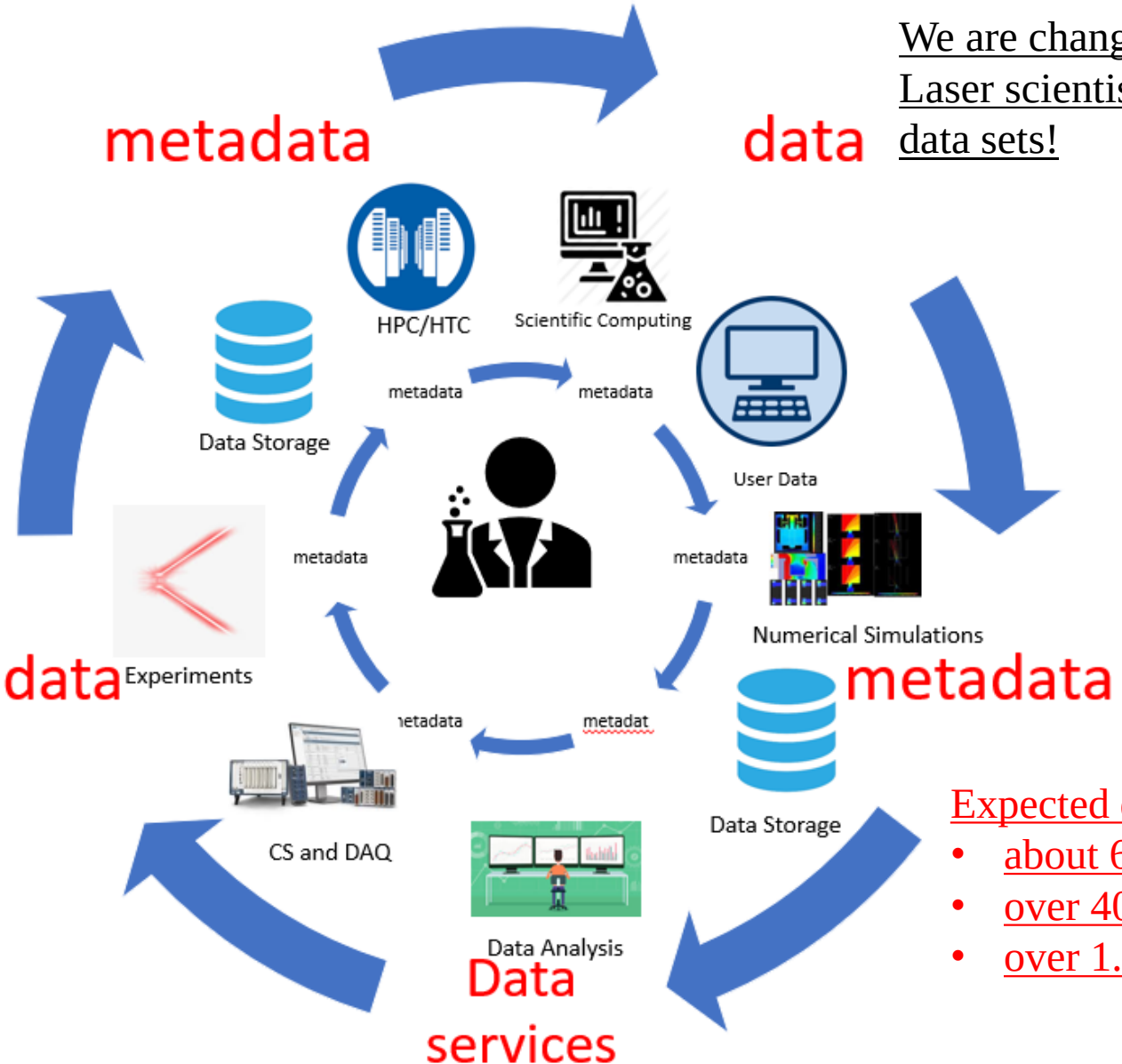
Conceptual work on

- 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 ELI-ERIC level

- Data policy significantly affects obligations ("FAIR")
- Metadata + UIDs + formats + .. affect current service development

ELI as a DATA Facility



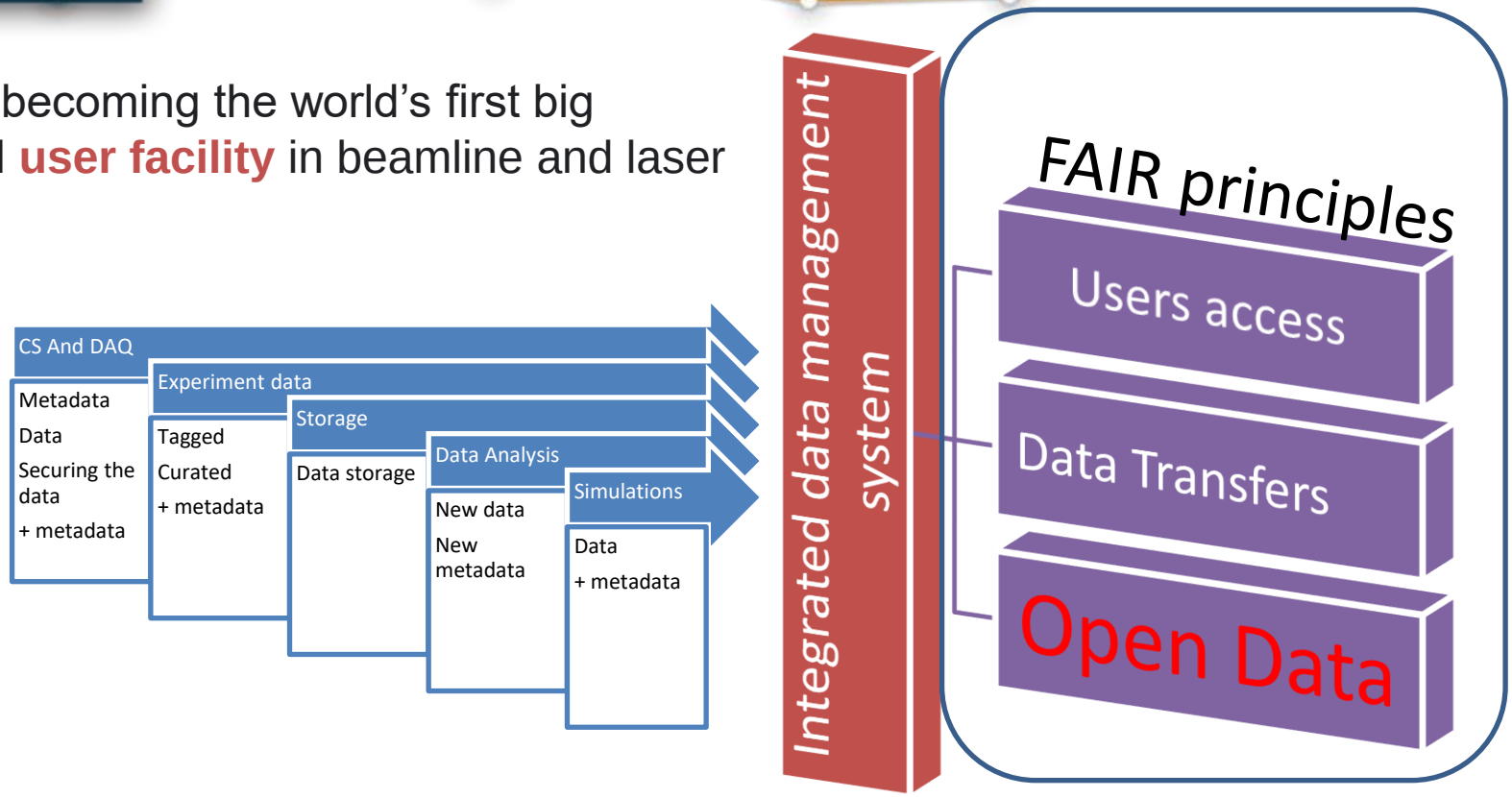
We are changing the mentality!!!
Laser scientists are not used with big data sets!

Expected data size 600 TB (1 month)

- about 64 000 USB drives or
- over 400 KG storage server or
- over 1.2 Mil DVDs

ELI FAIR Data Services

ELI aims at becoming the world's first big international **user facility** in beamline and laser research.



- **ELI's Integrated Scientific Data Management Systems will be the first of its kind.**
- **The design of our systems start with FAIR data by design, allowing us to be among the first facilities providing FAIR experiments. ELI has the chance to become one of the first "FAIR Facility".**

Open Data @ELI

Federated services – The ENGINE

The “abstraction layer” that allows the Data to be findable by the user:

- As part of the PaN community, ELI is currently adopting a wide range of services
  - Umbrella ID – Federated ID provider for the PaN community (EOSC integration)
  - Data Portal – for Open Data (federated-EOSC integration)
  - Federated Search API – as a microservice used by the Data Portal, making the Open Data searchable (EOSC integration)
  - REMOTE DATA Analysis is also considered, a service that aims to be supported via the Open Data Portal

Advantages:

- All services are scalable
- We can easily integrate data from our sites(ELI sites are in 3 different countries)
- Better understanding of our data

Disadvantages:

- New cost drivers have been identified
- New security challenges
- Computing limitations (hardware, software, skilled developers)

Open Data @ELI

Open Data – The Fuel

"***A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike.***" – opendefinition.org

ELI – Open Data will be published after the embargo period

- The PI, in the proposal, decides on the embargo period
- The PI, can make the data available even before the embargo period
- ELI Data "Retention" - ELI aims at keeping the Data for 10 Y and Metadata forever
  - NOT all the Open Data will be stored online, part of it might be stored on tape storage and made available "on-demand"
  - Metadata will be online

Advantages:

- Open Data allows the communities to work together and set new standards
- Input from different scientific communities
- Data Analysis tools and services will be implemented

Disadvantages:

- Storage becomes more complex
- New security challenges
- How can a facility track publications, we need to train users to use the DOIs

Open Data @ELI

Open Science – The Driver

"The goal is to turn data into information, and information into insight." – Carly Fiorina (former CEO HP)

https://www.nature.com/articles/s41559-020-1109-6

Open Data @ELI

Open Science – The Driver

"The goal is to turn data into information, and information into insight." – Carly Fiorina (former CEO HP)

Federated services Engine

- Federated services require clear data standards and format- HDF5
- Rich metadata is supporting the federated search tools – Nexus
- AAI – is federated – Umbrella ID

Open Data FUEL

- Data will be available after embargo period in HDF5 and searchable
- Data portal will be offering access to metadata and data sets
- Service level definitions and service level agreements are needed!!

Open Science Driver

- We can commit on providing a very limited computing capacity via the Data Portal!
- Tracking scientific publications will be challenging for Open Data!

There is one single challenge, "open" needs to be better defined! (Is Anonymous access considered Open Access?

Thank you