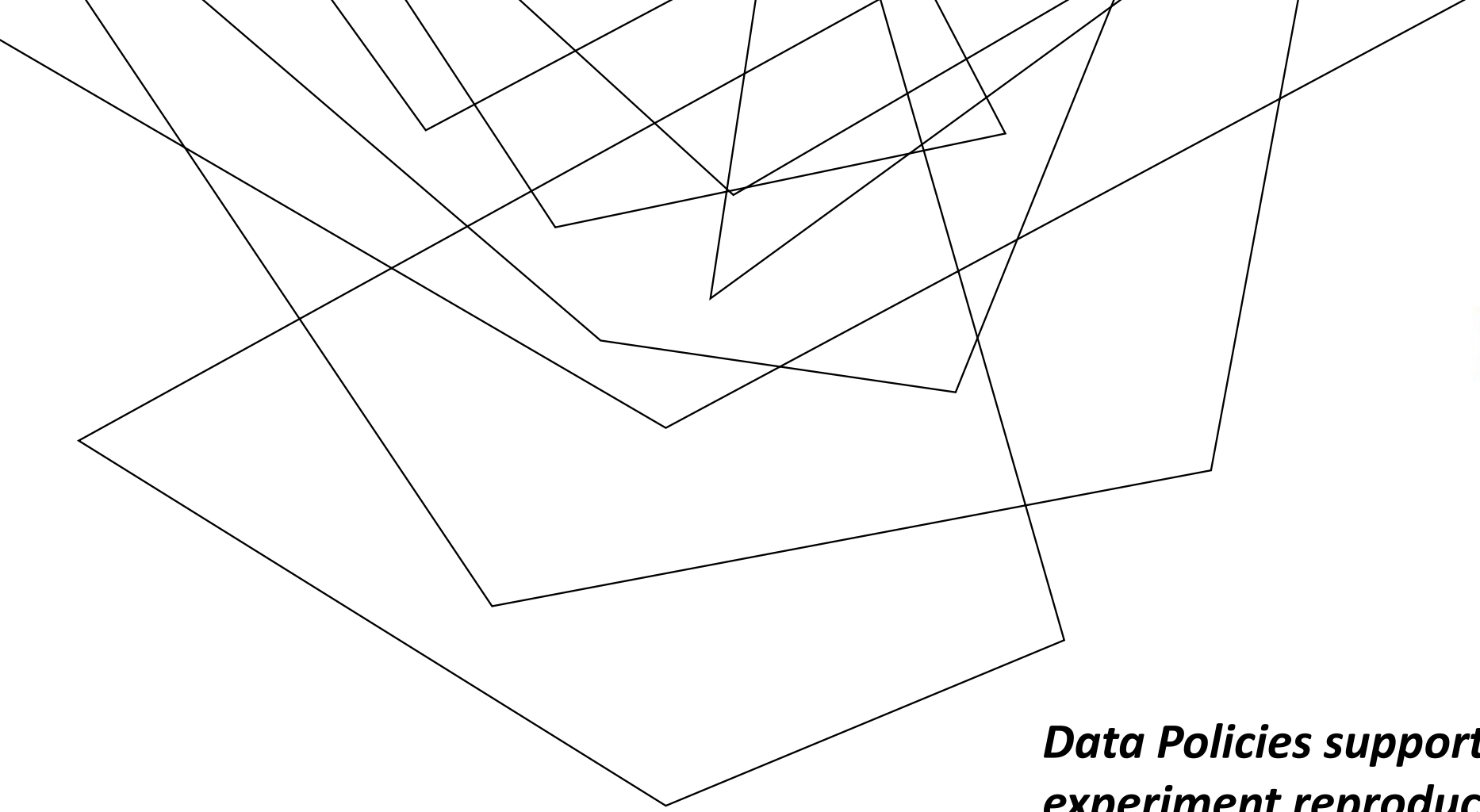


Data Policies supporting the FAIR standards/ experiment reproducibility challenges

Teodor Ivănoaica Lajos Schrettner

ELI ERIC – ELI ALPS

Overview

- Towards joint operation – ELI ERIC
- Integration of the pillars and integration into the community
- Data management developments

- Catalogue systems
- Logbook solutions
- Environment monitoring as metadata source
- Further challenges

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 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."

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."*

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

Common challenges are building communities!

What the users communities are asking:

- Good (meta)data + logbooks
- Performant Download services
- Digital Object Identifiers for Data
- Remote data analysis
- Access to Open Data
- Credit for Data re-use

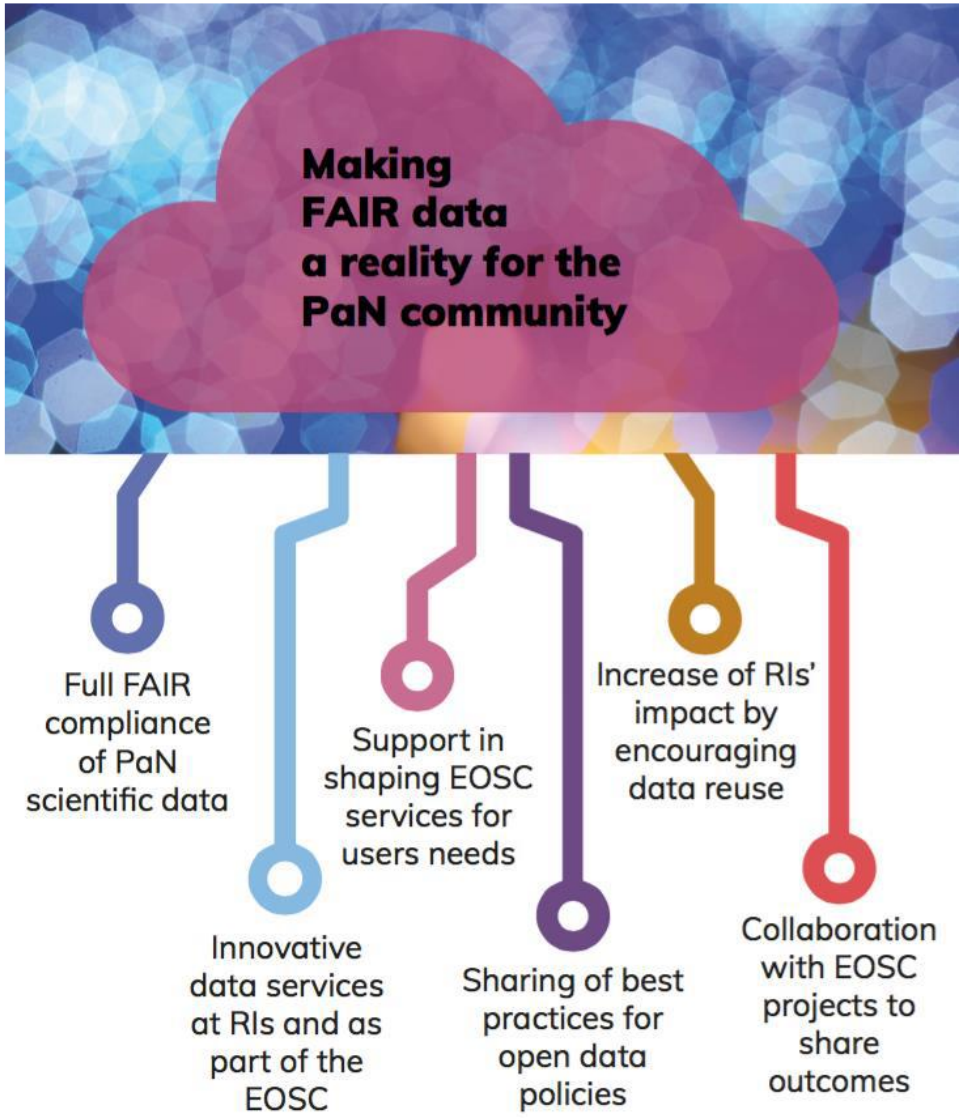
What Funders are asking:

- FAIR Data
- Open Science
- Digital Object Identifiers for Data
- Reproducible Publications
- Participate in the EOSC
- Metrics about Data Re-Use

What is happening pre/after discussing data?

- User: has an idea / need to study a sample
- Proposal: User writes a proposal for one of the facilities
- Beamline scientist: Review proposal and checks feasibility
- Review committee: Reviews proposal and rates scientific quality
- Beamtime allocated: User travels to facility / sends sample
- Experiment: Sample(s) are exposed to beam + data collected
- Analysis: Data is reduced, analysed+ curated (DOI)
- Publication: User publishes results (DOI) in peer review journal

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
  - ……

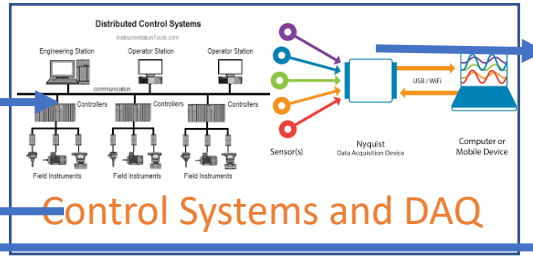
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852

Unified Data and Computing Experience

WHO?

Identity Management System?

WHERE FROM?



Control Systems and DAQ

Online storage

ELI ALPS/Beamlines/NP

HOW?

User portal

HOW TO LOOK FOR

Metadata

ICAT

SciCat

INVENIO

(META)DATA?

Ingest Metadata
Make it searchable

??offline and Long Term Storage??

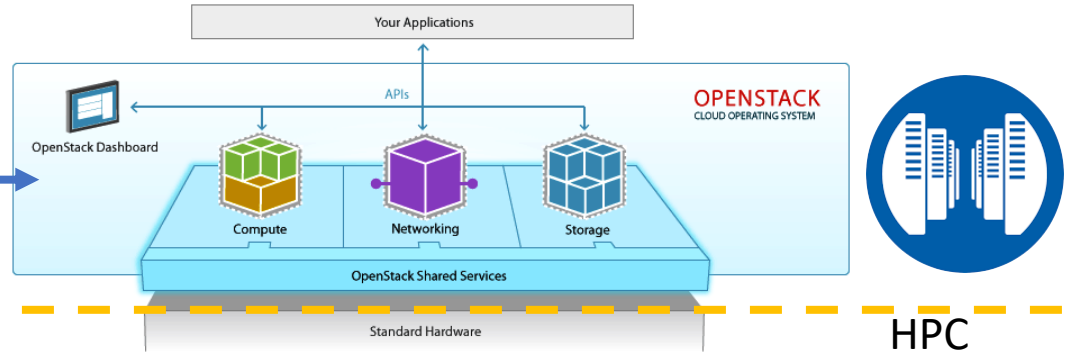
WHAT?

ELI internal (embargoed data)

panosc
photon and neutron open science cloud

ELI external portal for open data

WHERE TO ANALYSE?

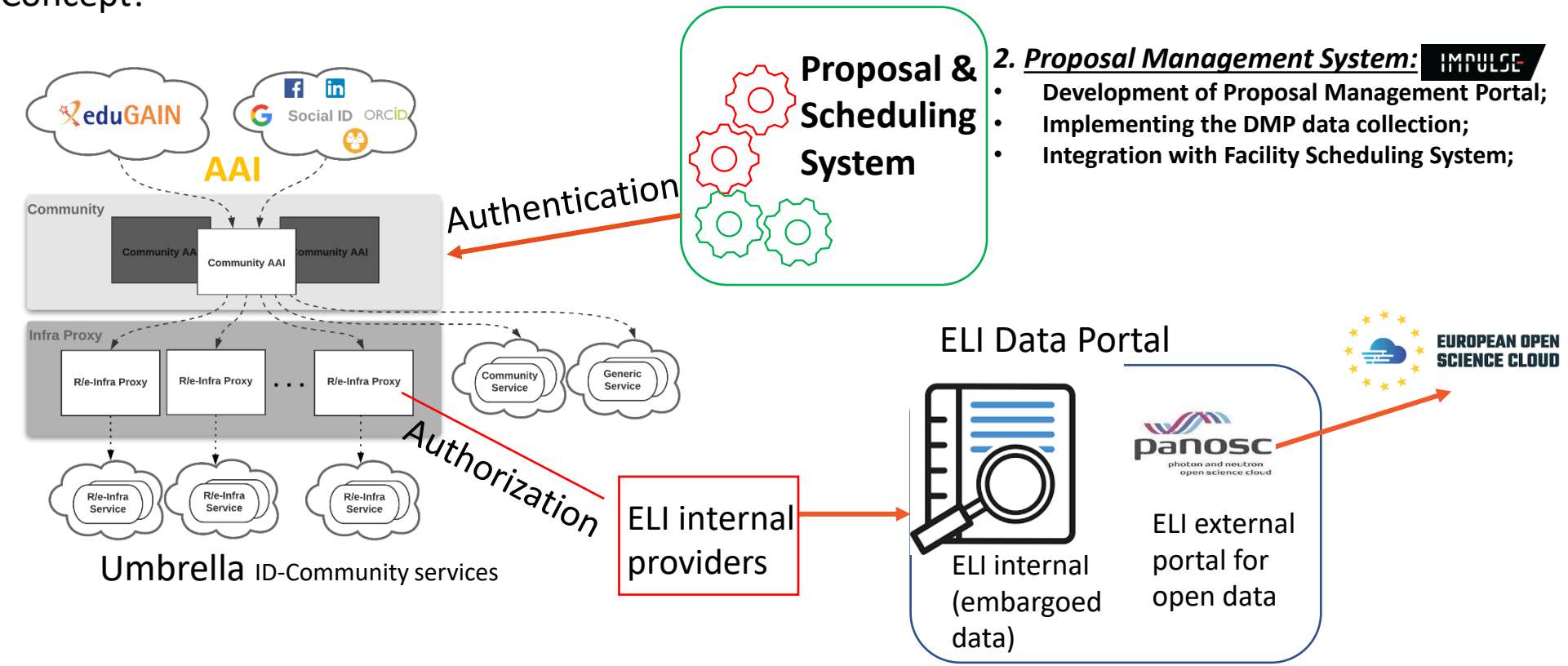


HPC

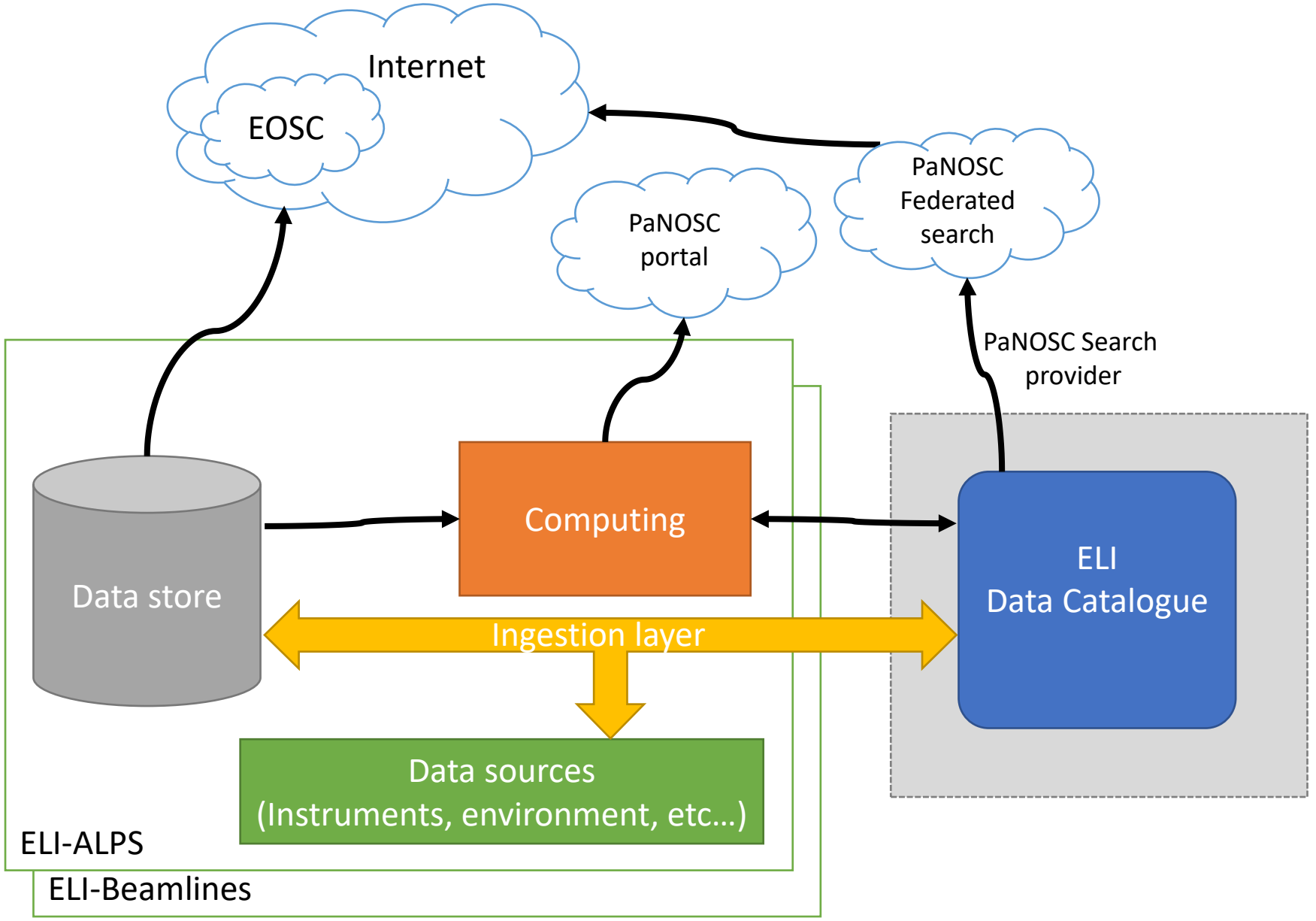
ELI Private Computing Cloud

Supporting the users by providing a fully integrated Scientific Data Management System!

Concept!



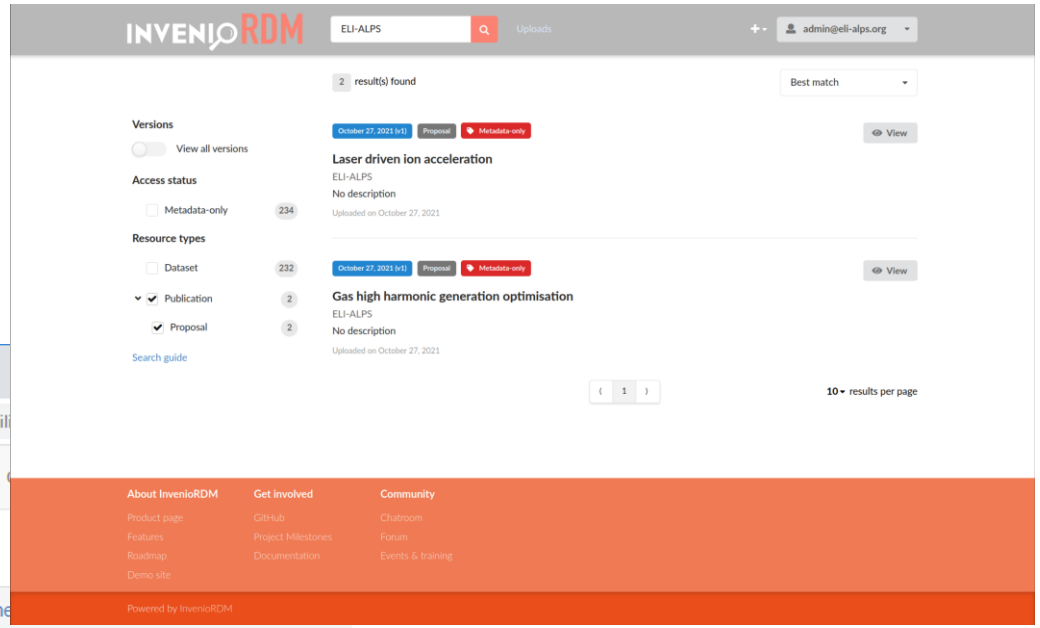
Data concepts and architecture

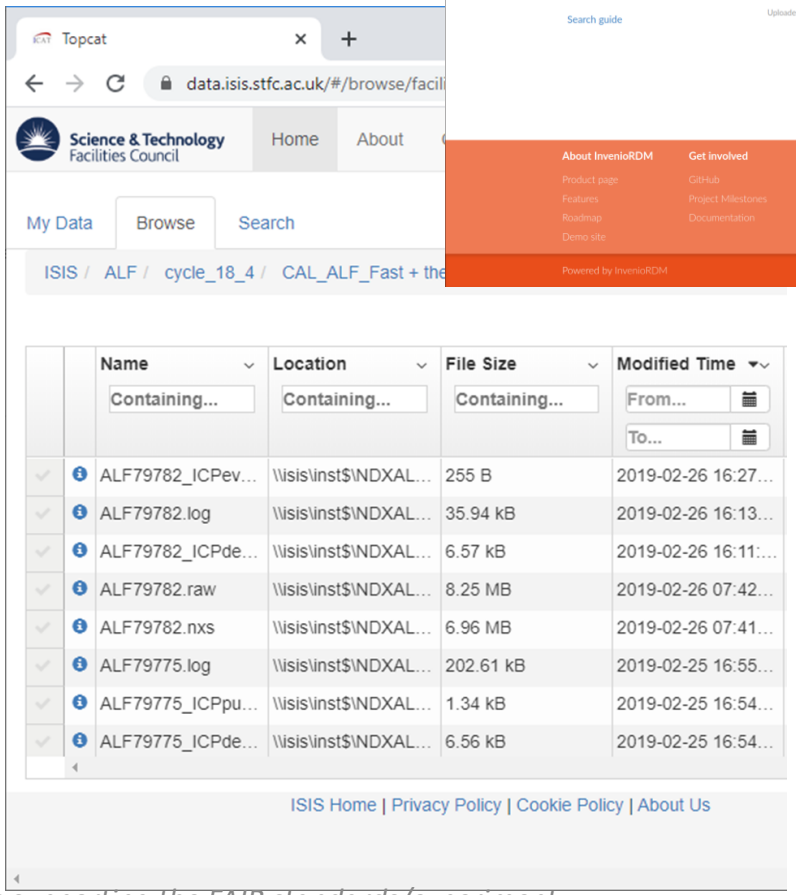


Catalogue systems evaluation

Summary

	ICat	SciCat	Invenio RDM
Pros	• used in PaN community • "PaNOSC ready" • „ERIC" support	• used in PaN community • "PaNOSC ready" • Compatible with modern solutions (containerization, kubernetes)	• Framework based • A customizable turn-key solution • Based on modern and exchangeable technologies/solutions • Good support • Commercial option
Cons	• EOL technologies (Angular JS, Python2) • Small developer community • Fragmented documentation	• Small developer community • „ERIC" support is not clear	• Under development – LTS in July 2021 • "PaNOSC unready"



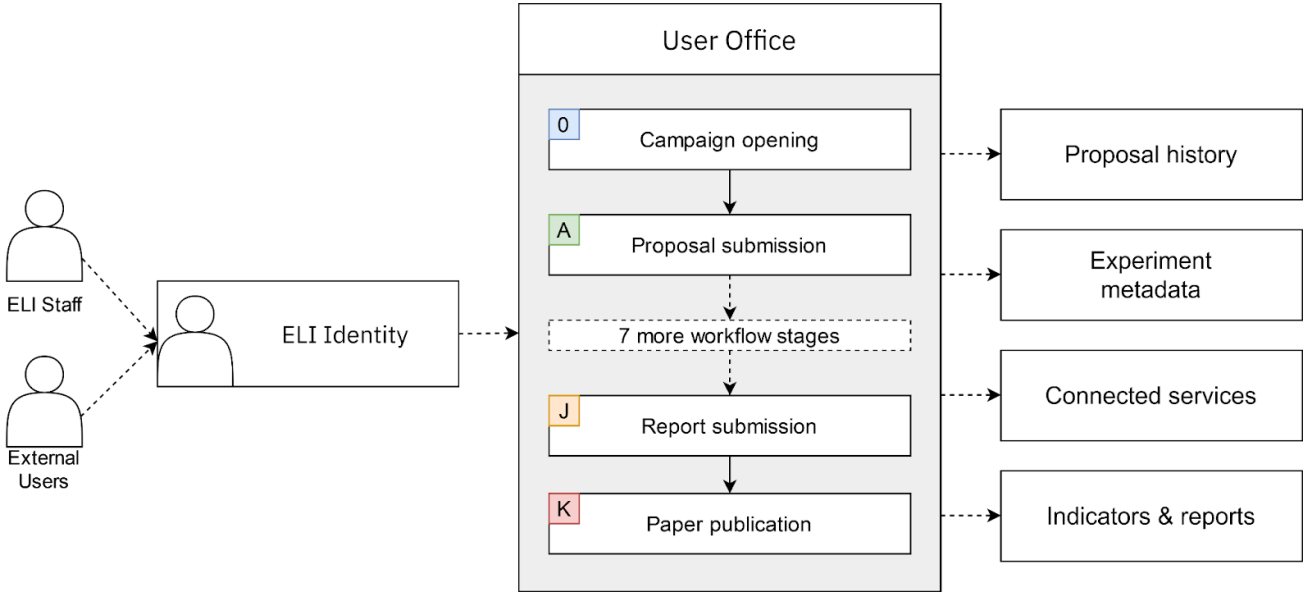


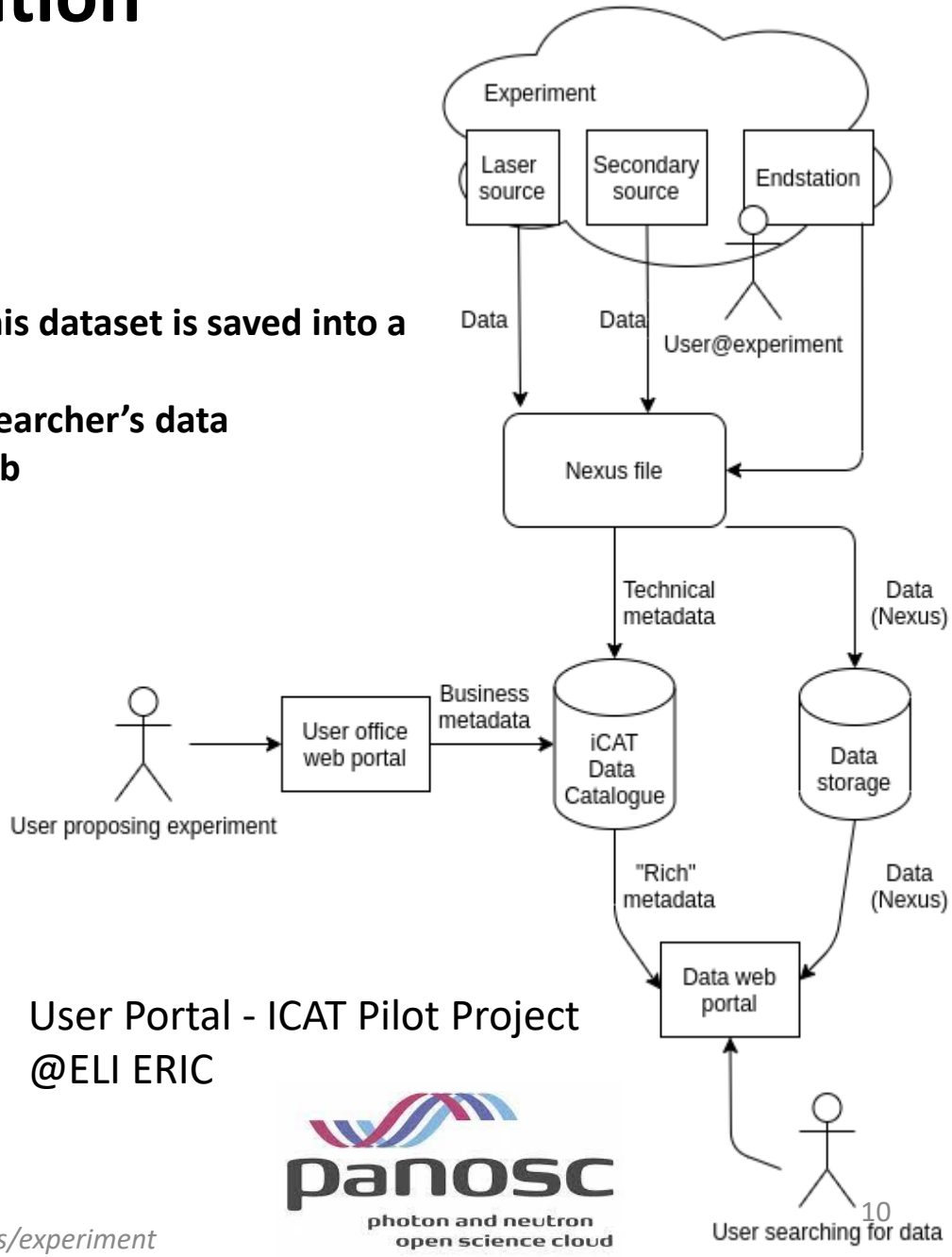
Catalogue systems evaluation

Pilots:

Catalogue integrations:

- **User creates proposal via User Office web portal**
  - **Defines what he wants to do, when and how**
- **User does the experiment, the full data adhering to his dataset is saved into a single Nexus file**
- **User wants to search for his old data/some other researcher's data**
  - **Searches and downloads via the Data Portal web**



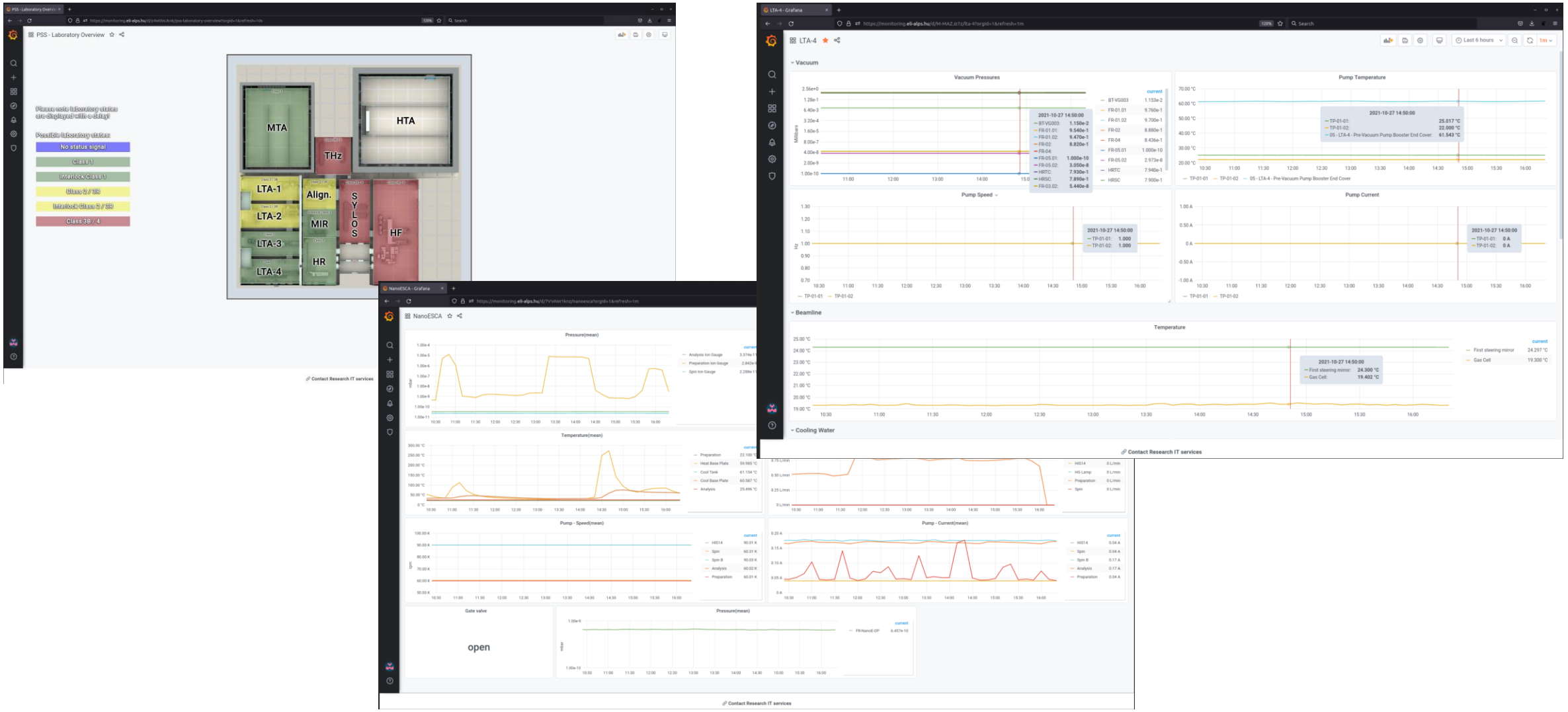


User Portal - ICAT Pilot Project @ELI ERIC

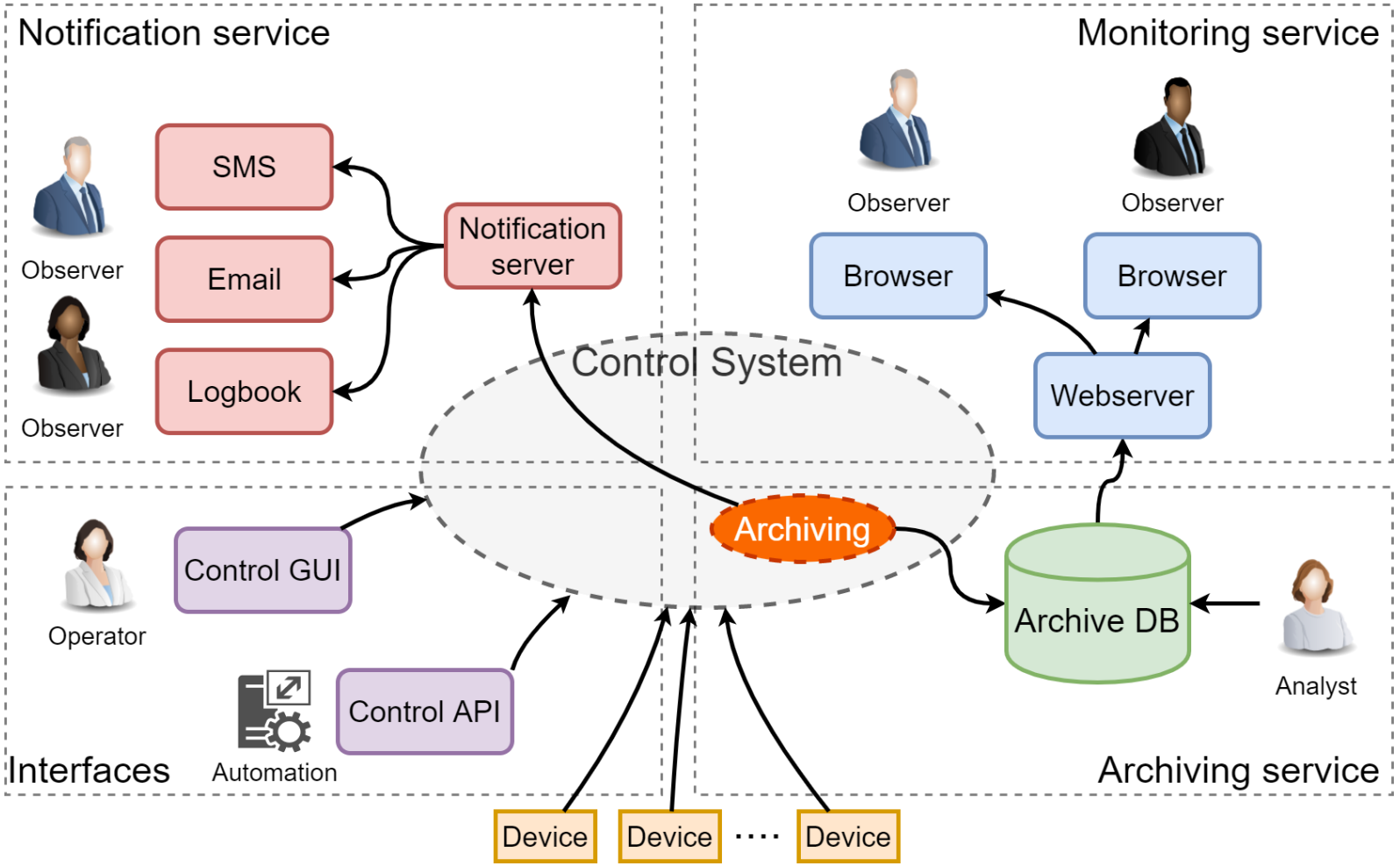
Data sources @ELI

- Proposal management system in development in IMPULSE
- Simulation capabilities per site -- HPCs
- Improve the raw data collection – input is needed
  - ALPS/Beamlines/ERIC – together to identify the possible solutions
- Post processing/analysis capabilities per site
  - Remote data analysis solutions considered (PaNOSC /VISA)
  - Other solutions (HPC) Jupyter Notebook-SLURM integrations
  - Needs/Suggestions are welcomed
- Metadata collection to be standardized/improved
  - Rich meta data strategy following FAIR standards addressing community needs
  - Electronic Logbooks/in-house as it was identified by our internal users (ELI ERIC-Beamlines-ALPS solution)

Customized dashboards facilitating operations

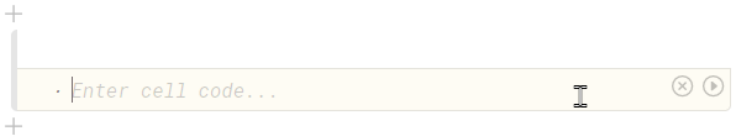


Environment data collection



Logbook evaluation

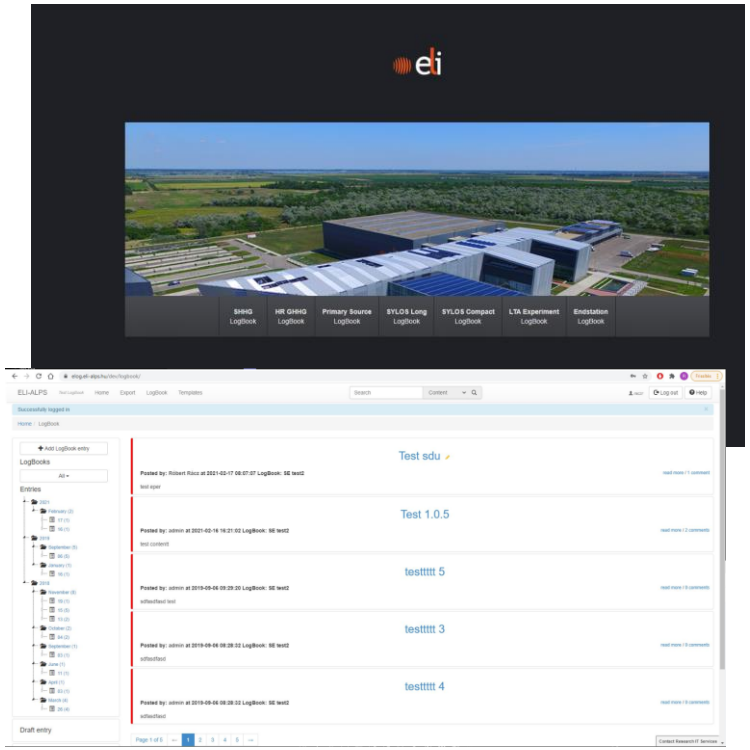
Distributed lookbooks technology

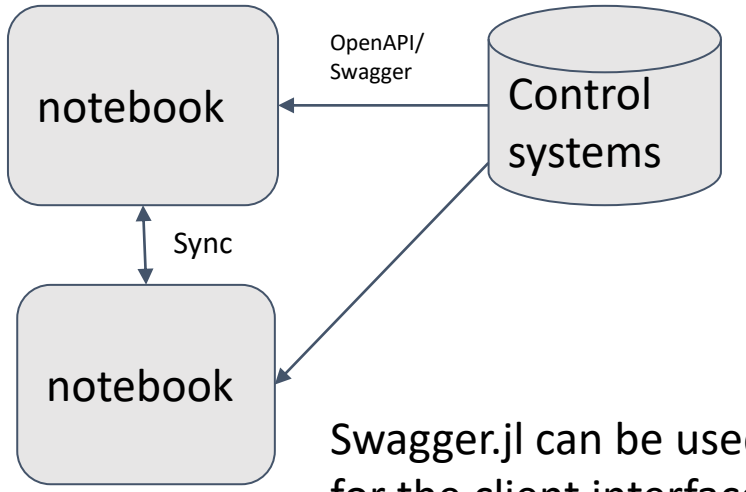


Pluto notebooks

- started with one PhD student is ready to be tested, another pilot to be started also @ELI ALPS
- Coherent notebook state for everyone
- Simple, easy to share
- Code will be documented and shared

ALPS Notebooks





Swagger.jl can be used to generate a Julia package for the client interface form a .json spec.

Integration challenges

- Site level
  - Heterogeneity of devices and scientific collaborations
  - Missing and/or partial interfaces of procured devices which usually require custom interfaces to be built
  - Standardization of data and data-related activities based on specific users requirements – application definitions using NeXus
- ELI wide
  - Integration of the DATA Systems and standards/identity management and federated services
  - Harmonisation efforts in IMPULSE supports the transition and integration of both management and scientific tools and services

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

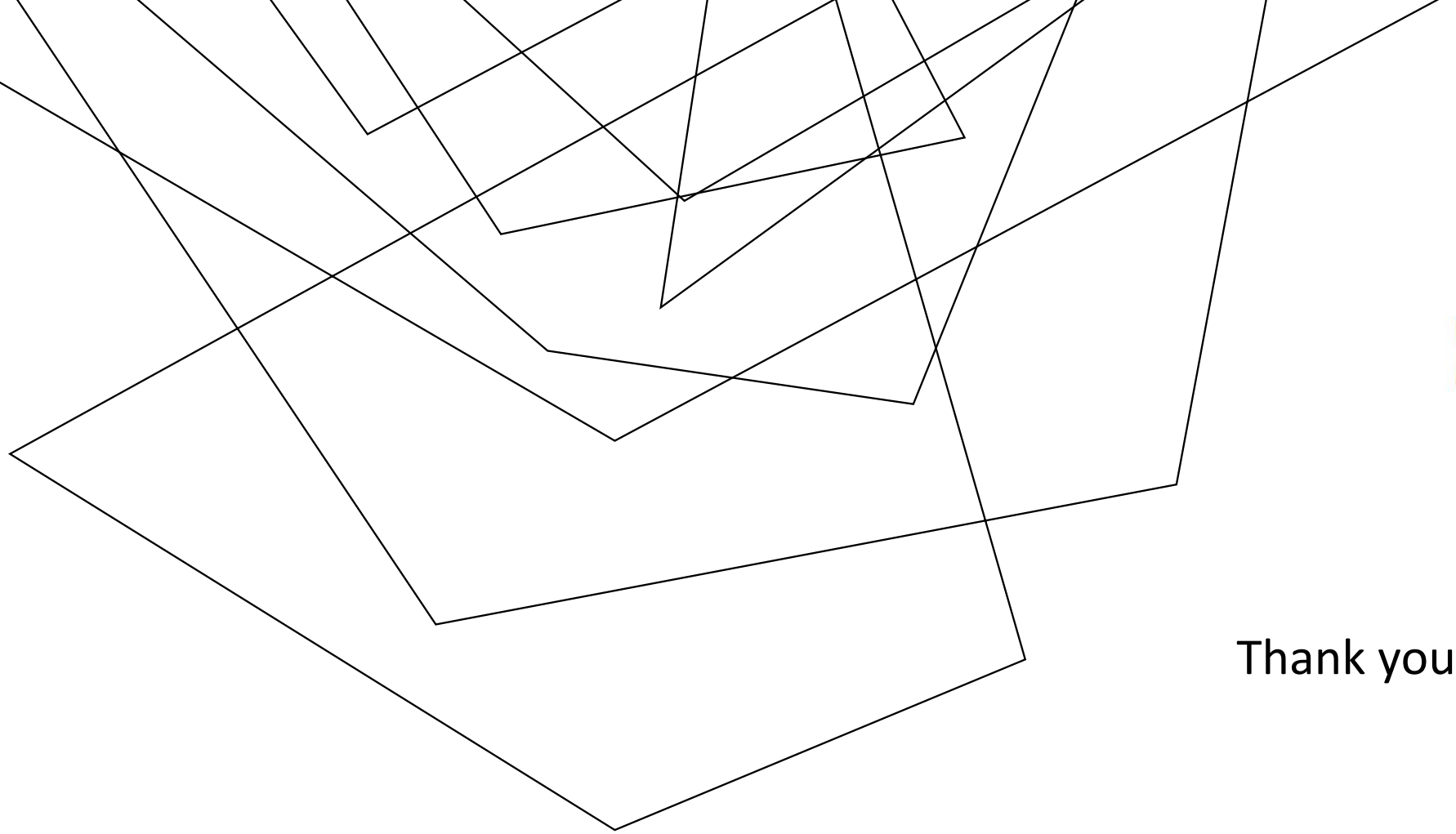
QUB
The Queen's University of Belfast

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.



Thank you for your attention!