RHIC Data and Analysis Preservation Round Table

05/08/2025

Introduction

Notes from previous meetings

- Notes from previous meetings are accessible on individual Indico pages.

- Please review the notes and provide comments if necessary.

Data and Analysis Preservation Timeline

Data and Analysis Preservation Timeline

End-of-Life for Core Infrastructure

- Compute (CPU), disk storage, and tape I/O systems
- Existing capacity will not be maintained

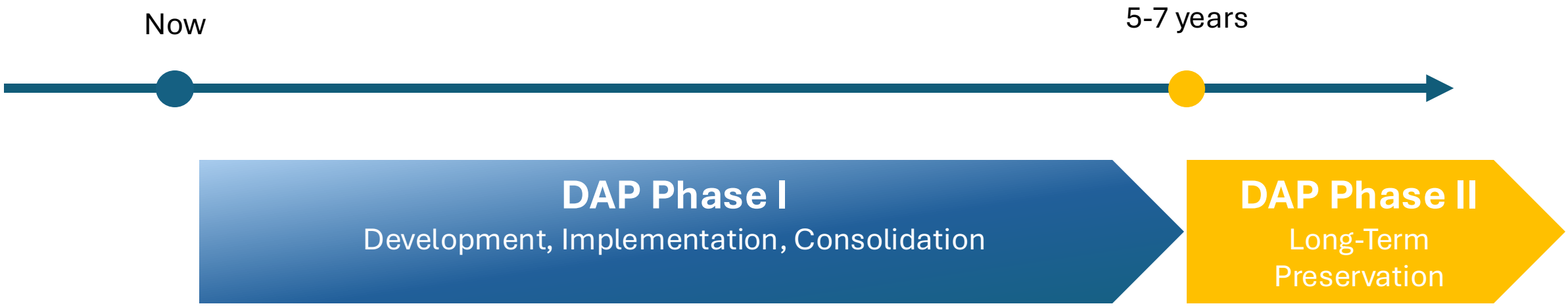
Limited Future Processing Capability

- Reprocessing a sizable data sample within a reasonable time frame will become increasingly difficult

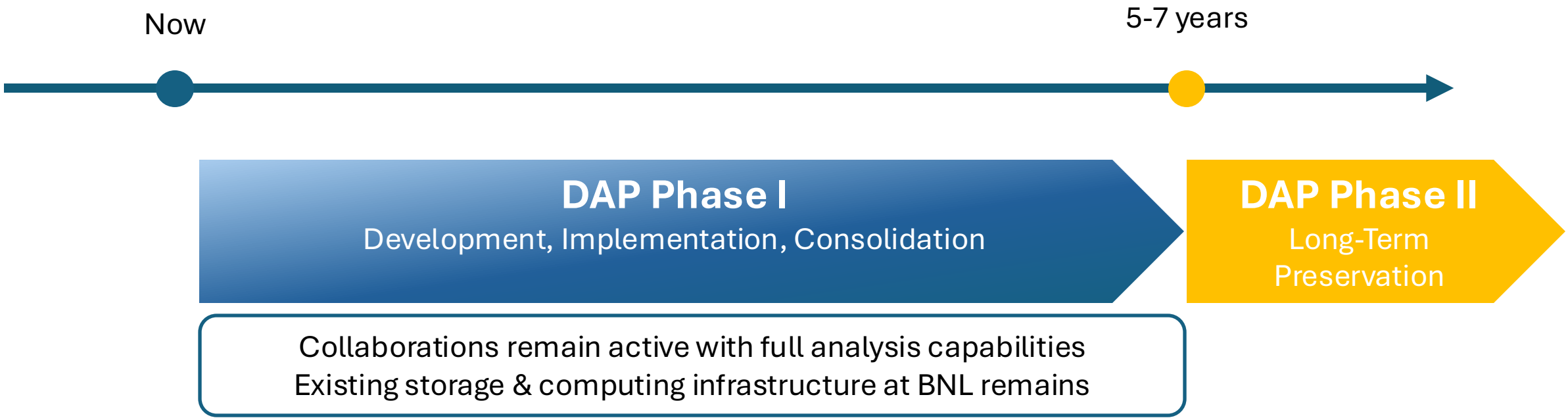
Limited analysis capability?

Long-term Data Preservation Starts

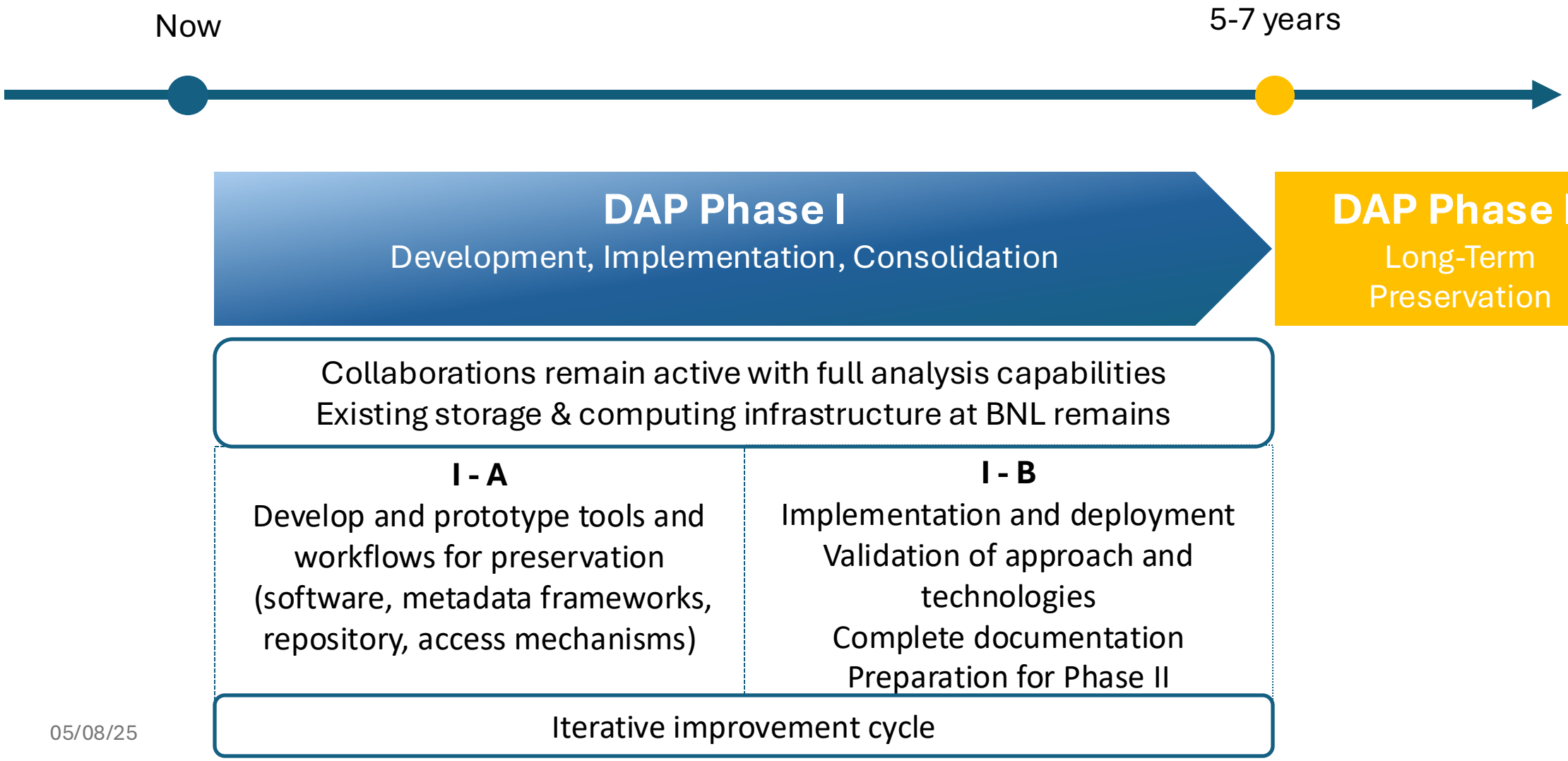
Data and Analysis Preservation Timeline



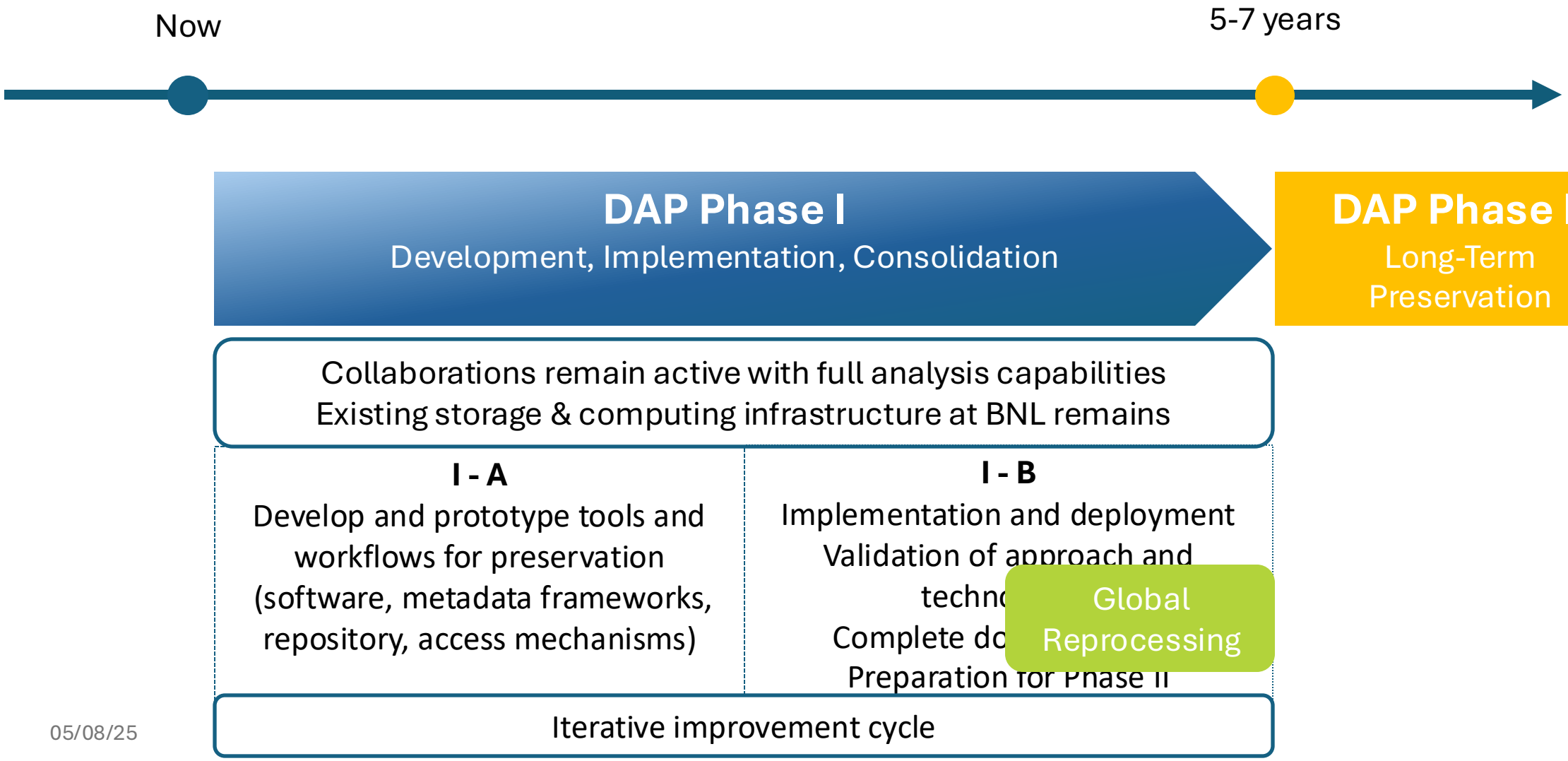
Data and Analysis Preservation Timeline



Data and Analysis Preservation Timeline



Data and Analysis Preservation Timeline



A Global RHIC Data Reprocessing

Comprehensive final reprocessing of data before bulk compute and tape bandwidth are reduced

Uses uniform validated software versions to create a consistent baseline dataset for future research

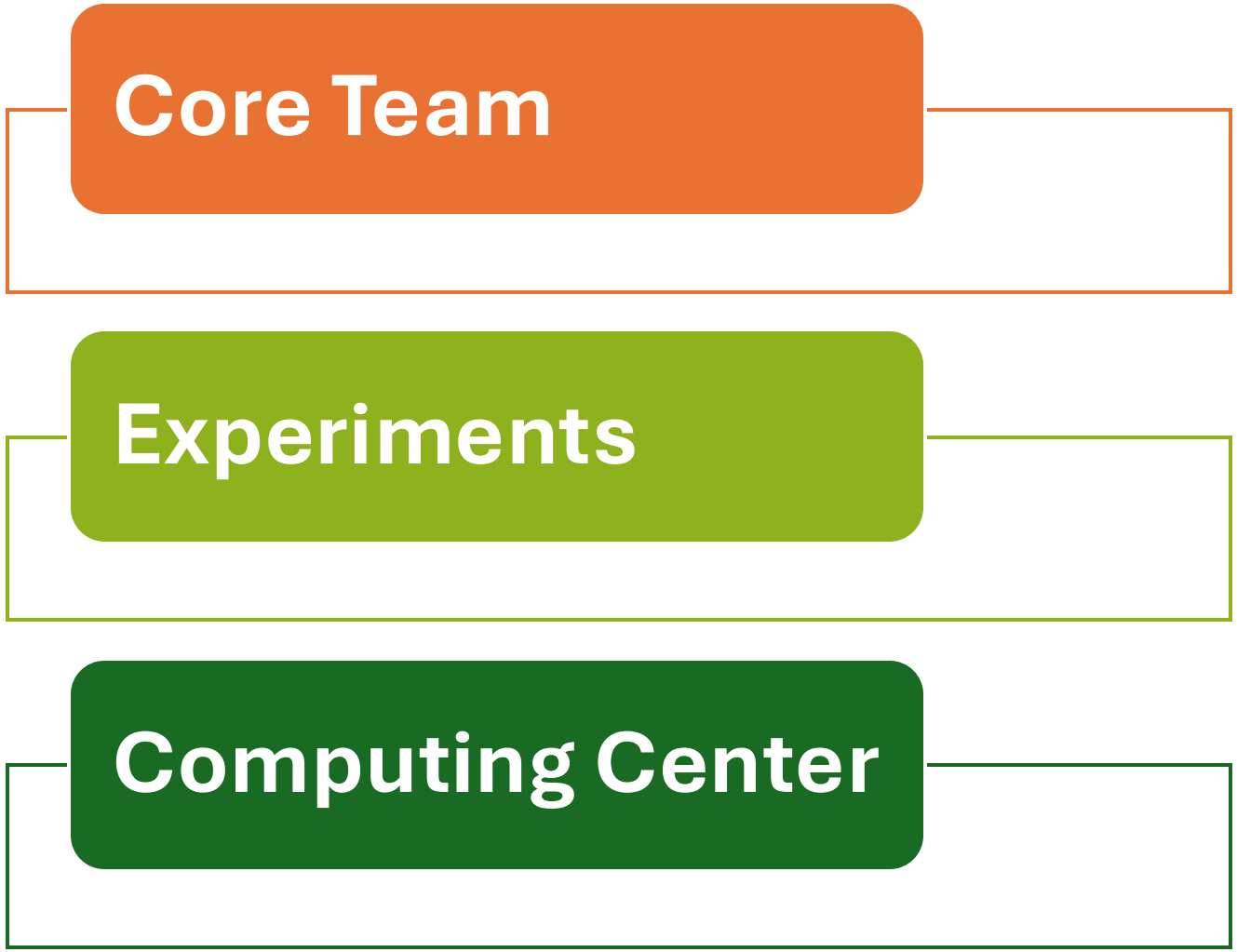
Unified Dataset
Consistent calibration, reconstruction algorithms, and data structures across all run periods

Preserved Collaboration Knowledge
Captures expertise while resources remain available, incorporating refined understanding of detector performance and physics processes

Documented Processing Pipeline
Complete workflow from raw data to analysis objects, serving as documentation and validation

Simplified Provenance Tracking
Single, definitive version of reconstructed data rather than multiple historical processing versions

RHIC Data Preservation Staffing Model

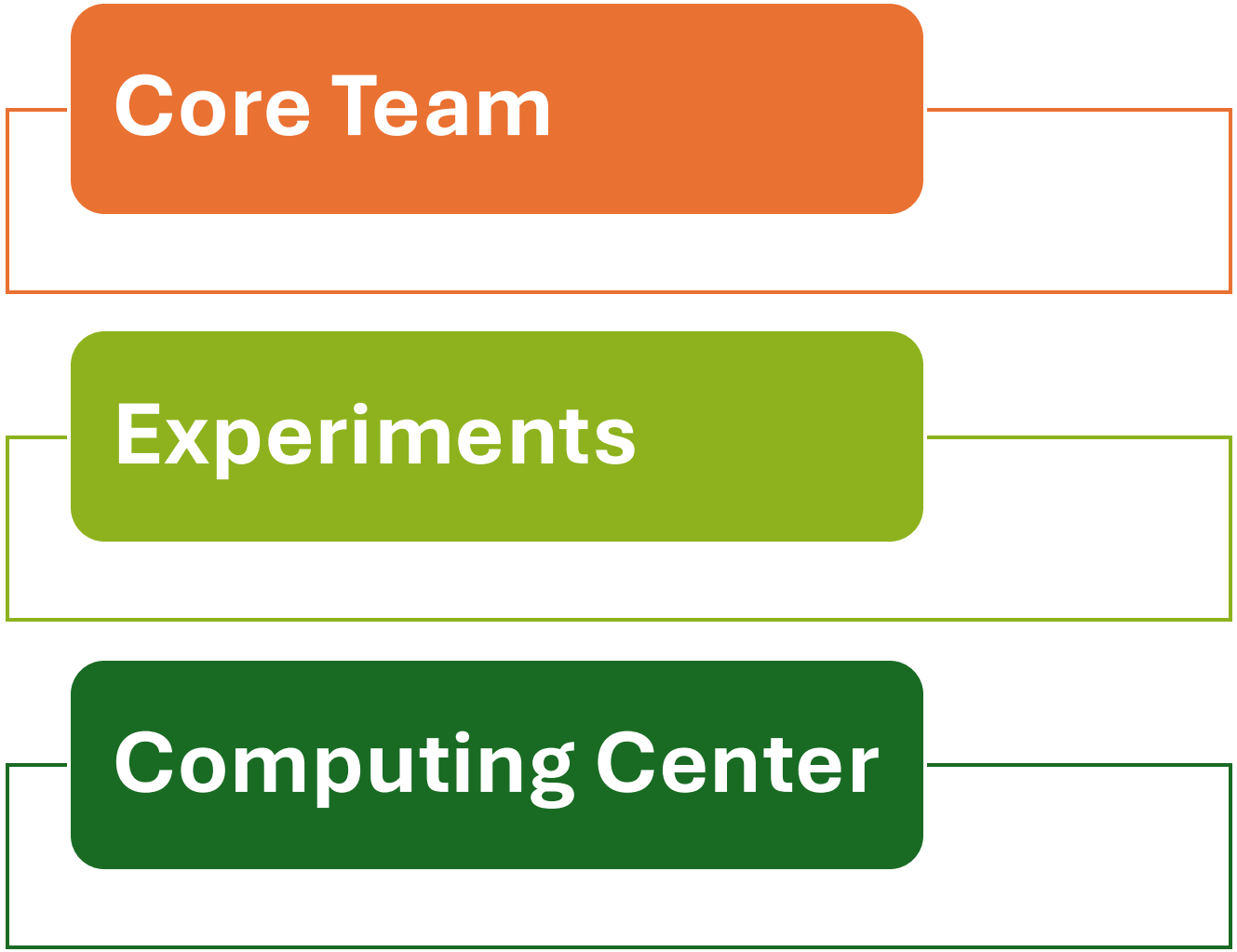


Community Alignment

- Align with national & international efforts
- Apply FAIR principles
- Support shared metadata & data access
- Coordinate with NP community
- Ensure interoperability with standards

Team Resilience

- Cross-train team members
- Avoid single points of failure
- Ensure smooth staff transitions
- Share knowledge regularly
- Keep documentation up to date



RHIC Data Preservation Roles

Core Team

- DAP Manager
- Software & Workflow
- Repository Systems
- AI Integration
- Web Development
- Documentation & QA

Experiments

- Representatives
- User learning and training

Computing Center

- Computing Center Liaison
- User Support
- Technology Watch

RHIC Data Preservation Roles

Core Team

DAP Manager
- Leads overall preservation strategy
- Coordinates team activities
- Manages stakeholder relationships
- Ensures alignment with DAPP and institutional goals

Software & Workflow
- Support for analysis and software preservation
- Maintains containerized environments, VMs, and workflow systems such as REANA
- Builds and documents reproducible computing environments

Repository Systems
- Manages Invenio-based digital repository
- Integrates with metadata and storage systems
- Supports OpenData and cold storage coordination

AI Integration
- Connects with AI/ML development groups
- Integrates AI for metadata, search, and user interfaces
- Aligns tools with RHIC data structures and workflows

Web Development
- Builds user-friendly simple static web interfaces
- Ensures accessibility and responsive design

Documentation & QA
- Maintains process documentation
- Conducts quality checks and standards compliance

Experiments

Computing Center

RHIC Data Preservation Roles

Core Team

Experiments

- **Representatives**
  - Ensure preservation aligns with PHENIX, sPHENIX, and STAR needs
  - Embedded in collaborations to connect scientific knowledge with preservation strategies
  - Identify, validate, and organize key datasets and documents for long-term use
  - Ensure comprehensive metadata, documentation, and scientific context
- **User Learning & Training**
  - Designs training programs and online resources to support user onboarding and effective reuse of preserved data

Computing Center

RHIC Data Preservation Roles

Core Team

Experiments

Computing Center

Data Access and User Support

- Supports researchers accessing preserved data
- Helps new users beyond the original collaborations

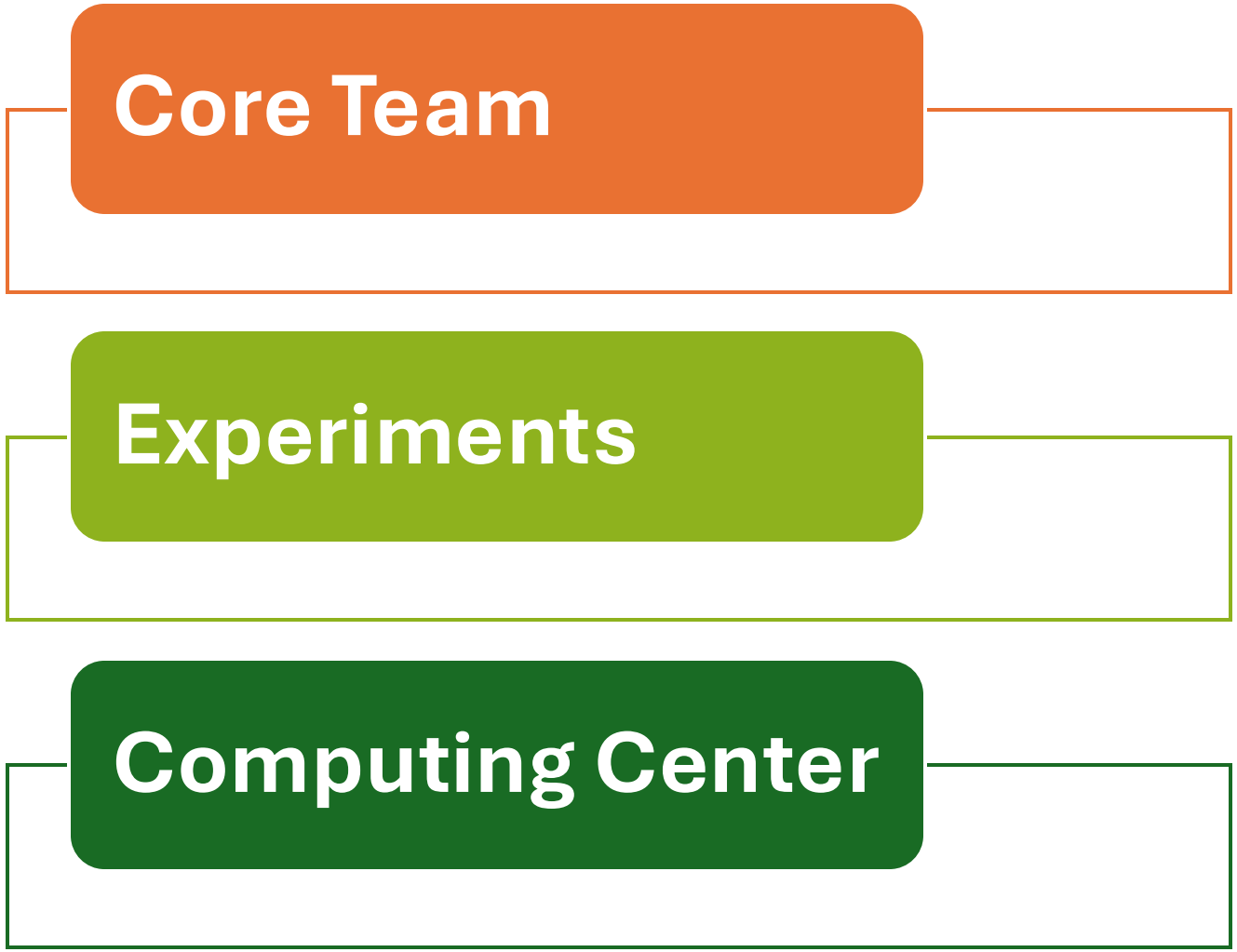
Technology Watch Analyst

- Tracks emerging technologies and obsolescence risks
- Recommends updates and migrations
- Produces an annual technology report

Computing Center Liaison

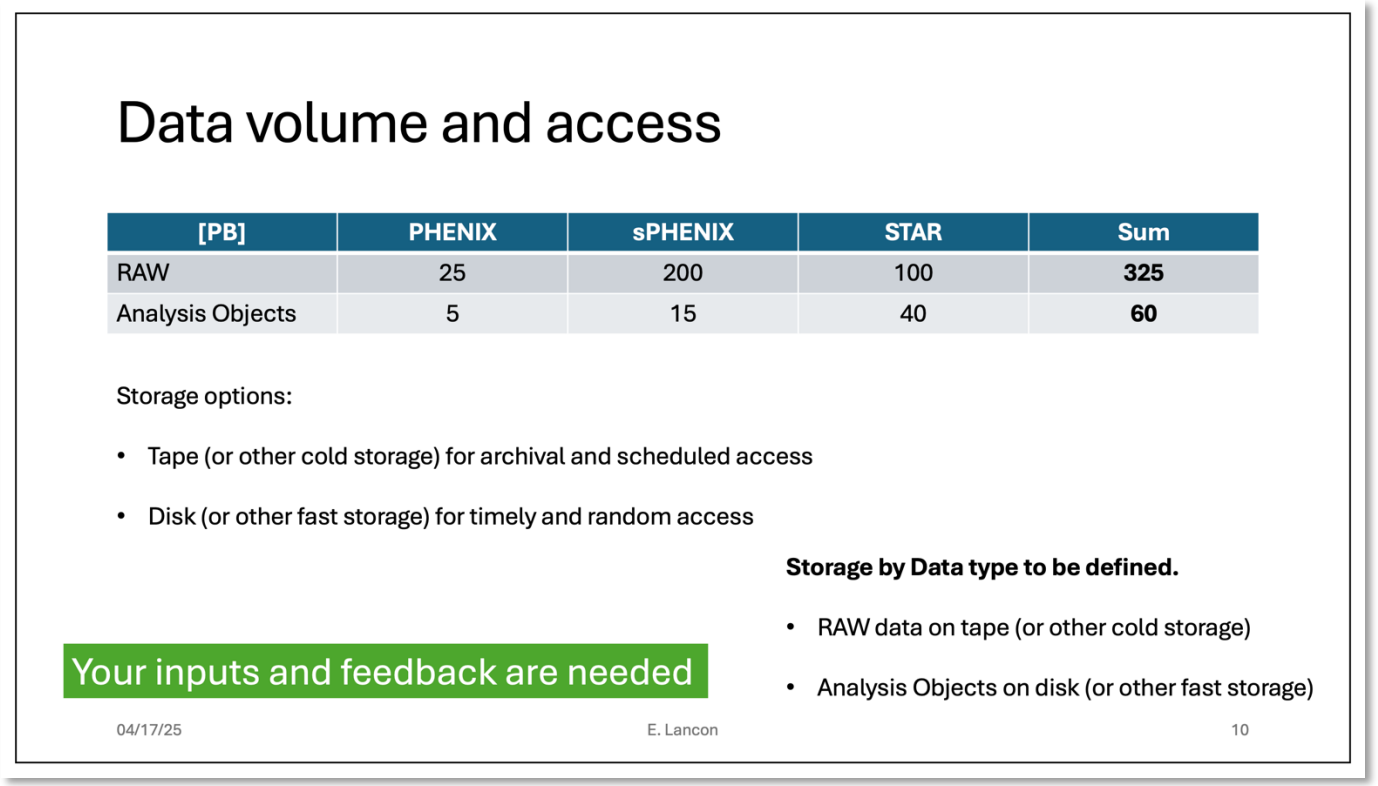
- Coordinates between the DAP team and computing center staff
- Ensures integration with infrastructure and ongoing support

Your feedback on the proposed roles is needed



Which data volume needs to be preserved?

- Which data need to be preserved?
- Inputs are required to guide the DAP implementation
- **Thank you for your feedback!**



Data volume and access

[PB]	PHENIX	sPHENIX	STAR	Sum
RAW	25	200	100	**325**
Analysis Objects	5	15	40	**60**

Storage options:

- Tape (or other cold storage) for archival and scheduled access
- Disk (or other fast storage) for timely and random access

Storage by Data type to be defined.

- RAW data on tape (or other cold storage)
- Analysis Objects on disk (or other fast storage)

Your inputs and feedback are needed

04/17/25 E. Lancon 10

Data volume and preservation levels

[PB]	PHENIX	sPHENIX	STAR
RAW	20	160-300	100?
Analysis Objects	5	50-100 (one processing)	40?
Other archive	10	50-100 (prev. processing)	?

Other data: historical data, previous processing, etc..

Need to be added for planning purpose

Data volume and preservation levels

[PB]	PHENIX	sPHENIX	STAR
RAW	20	160-300	100?
Analysis Objects	5	50-100 (one processing)	40?
Other archive	10	50-100 (prev. processing)	?

[PB]	PHENIX	sPHENIX	STAR	Sum
Level 3 (AO only)	5	50-100	40?	**95 – 145?**
Level 3' (AO + archive)	15	100-200	?	
Level 4 (RAW + AO + archive)	35	260 - 500	140?	**435 – 675?**

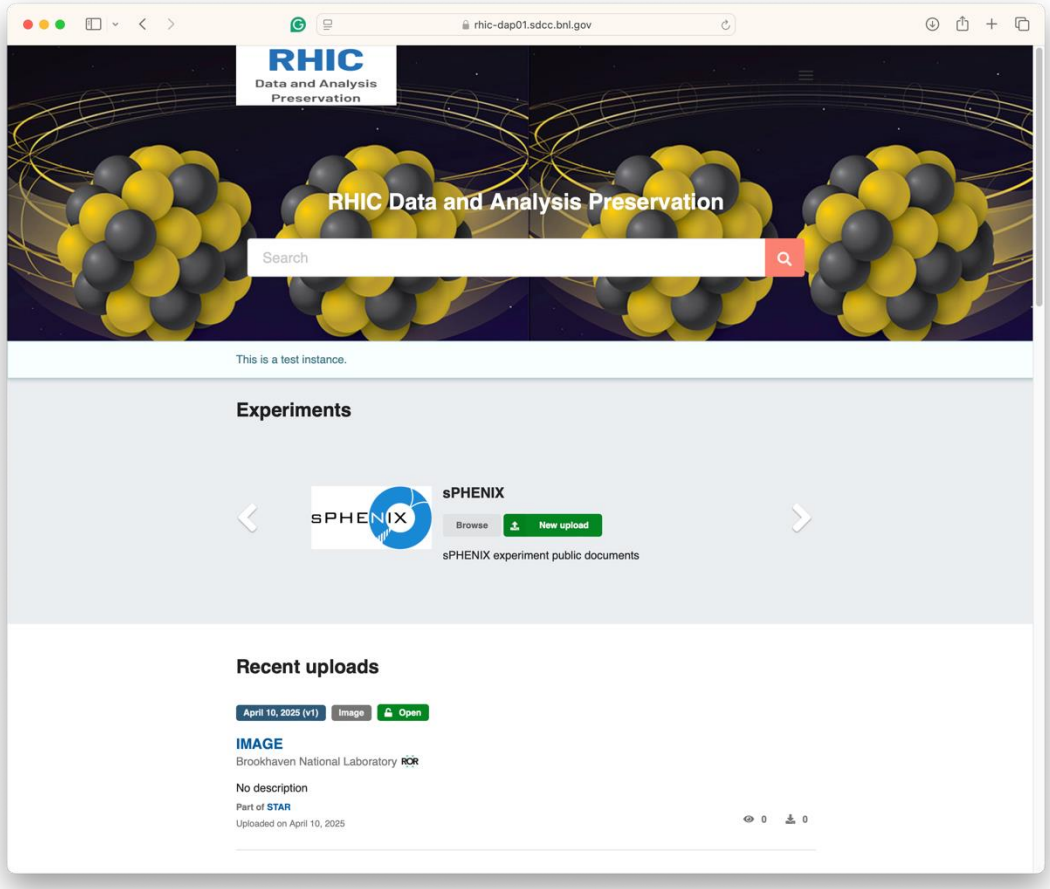
Level 3 on disk will be challenging
Exercise to be continued.

Reana status update

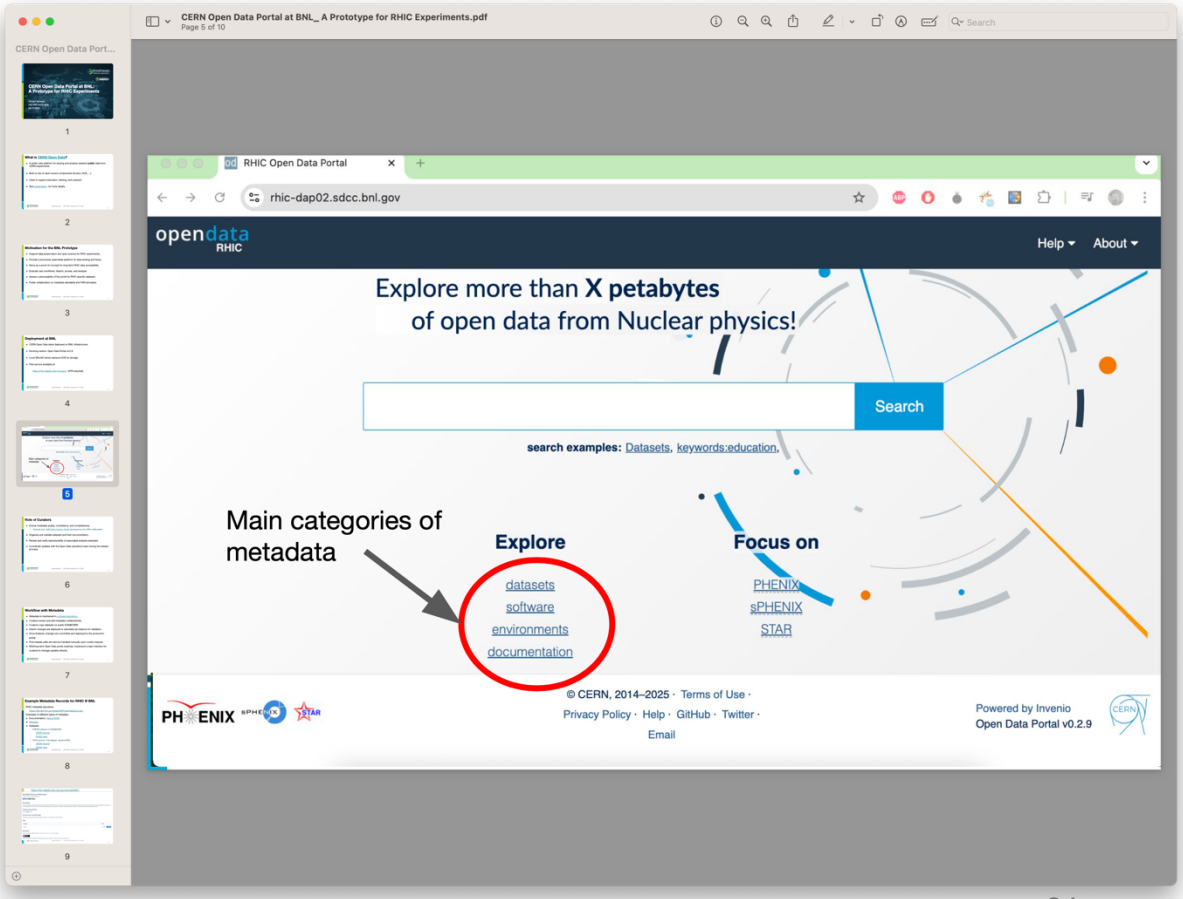
- The current PHENIX REANA instance is unmaintained, outdated, and lacks proper documentation; it is scheduled for retirement.
- **New Instance**: A limited-size replacement on OpenShift has been part of the DAP project plan from the beginning, with effort identified.
- **Current Status**: OpenShift infrastructure has stabilized. The deployment of the new instance is pending hardware allocation from the computing center.
- Once hardware allocation is complete, proceed with the deployment of the OpenShift-based replacement.

Give it a try!

https://rhic-dap01.sdcc.bnl.gov/ — BNL network access required



https://rhic-dap02.sdcc.bnl.gov/ (VPN required)



Actions from previous meetings

- Clarify whether the PHENIX taxi system needs to be maintained to preserve its analysis capability and what keeping it would entail.
- List of external dependencies
- Technologies for a Web repository, what are the replacements for Drupal
- Finalize data volume associated with Level 3 and Level 4
  - Experiments are requested to provide volumes of data to be preserved for the various categories
  - Review of databases needed for Level 3 and Level 4 preservation, respectively.
- Feedback from the experiments on BNL InvenioRDM and Open Data Portal instances
- RHIC Data Management Plan document to be circulated.
- Contact BNL's library about the release of OSTI's DOIs
  - BNL Library has been informed

Today

1. Update on OpenData at BNL - Vincent
2. Lightweight static web – Maxim
3. Hardware projections - Shigeki

- Next meeting: **Thursday, 15/08**