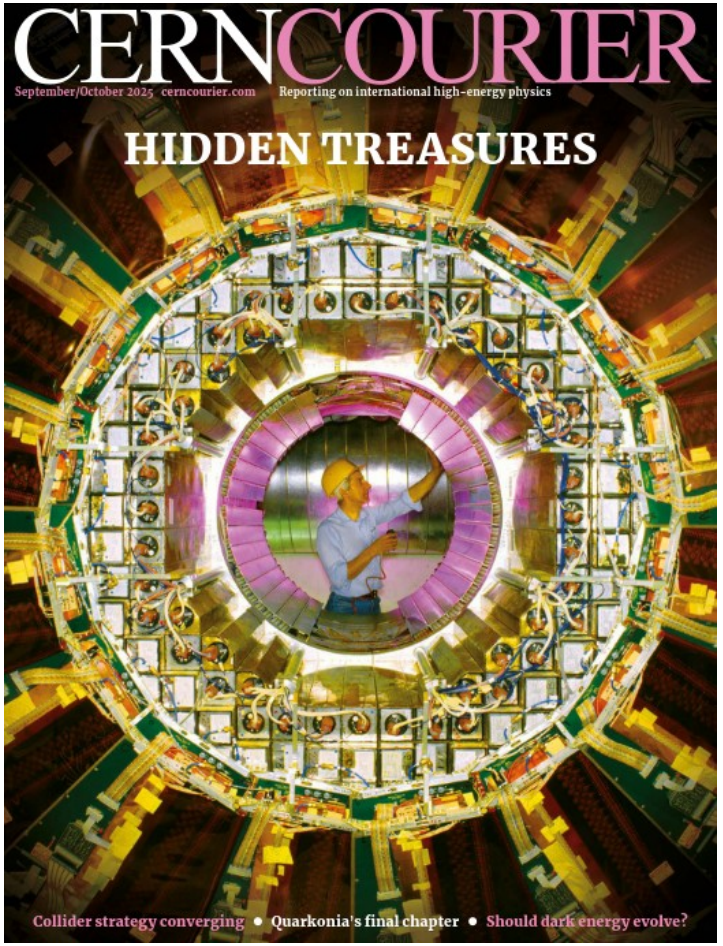
Secure data access to ancient data sets: XrootD in and for data preservation

Presenter: Ulrich Schwickerath
Date: 16 Sept 2026

Outline

- About Data Preservation in HEP:
  - What it is about
  - Why we do it
- DPHEP collaboration
- XrootD for data access
  - CERNLIB, PAW, Experiments
  -



About Data Preservation in High Energy Physics (DPHEP)

What is data preservation

“(...)The main goal of data preservation is to protect data from being lost or destroyed and to contribute to the reuse and progression of the data.”

Source: https://en.wikipedia.org/wiki/Data_preservation

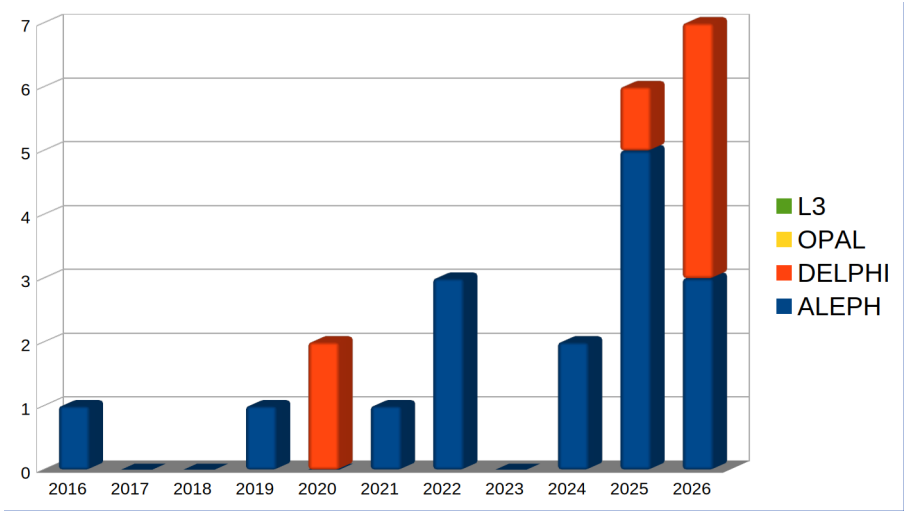
Data preservation: why?

- Unique data sets
  - Keeping them is **cheaper than recreating** them
  - In some cases they cannot be restored (e.g. meteorological data)
- Some possible **reasons to go back** to old data
  - New **analysis ideas**, e.g. new theories
  - **Advancements in theory**
  - Advancements in **analysis techniques**
  - **Learn** from old data for new experiments
  - ...
- Seen examples for each of these during 2025!
  - Older example: Recovery of JADE data, now available from https://opendata.cern
  - Mentioning this because it raised awareness of the need!

Data preservation: why?

Example LEP data:

- Increased interest in LEP data
- Mostly in the context of FCC
  - Migration to EDM4HEP format
  - New event generators
  - New techniques (e.g. jet flavor tagging)



Areas of data preservation

Bit - Preservation

Maintain and keep accessible:

- Raw data
- Reconstructed data
- Simulated data
- Meta data
- databases, conditions data

Software preservation

Includes:

- Analysis framework
- Reconstruction software
- Simulation software
- Visualisation

Note: Keep the sources (alive). Binaries may stop working

Documentation

- Published results
- Notes and internal documentation
- Technical documentation
- Manuals
- Web pages
- Meta data
- ….

Analysis preservation

- What has been done ?
- How has it been done ?
- Which data sets were used ?
- Analysis software sources

Ability to reproduce published results long term

Areas of data preservation: risks

Bit - Preservation

- Physical media failure (e.g. tape damage)
- System errors
- Human errors
- External events, disasters

Mitigation: dual-tape, external copies

Software preservation

- Ever changing IT infrastructure
- Architecture and technology changes
- Storage systems access changes
- Security enhancements, e.g. deprecation of old ciphers
- Compiler and computing language changes
- Deprecation of required external dependencies

Documentation

- Quality of internal documentation
  - May not be intended for general public
  - Use of jargon
- Author agreements, copyright
- Paper copies with bad quality
- Document format long term support

Mitigation: QC from the start of the experiment

Analysis preservation

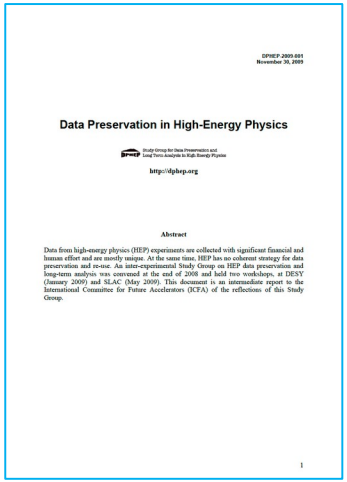
- Code not being provided and lost
- Code not maintained long term
- Proper documentation of the code and the analysis details

Mitigation: Guidelines for analyses, enforce these if possible

The DPHEP Collaboration

2009
Letter of intent

arXiv:0912.02


DPHEP-2009-001
November 30, 2009

Data Preservation in High-Energy Physics

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

http://dphep.org

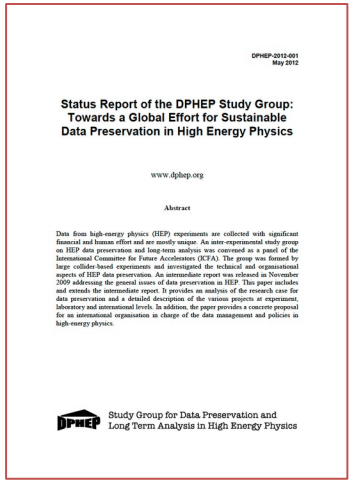
Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. At the same time, HEP has no coherent strategy for data preservation and re-use. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened at the end of 2008 and held two workshops, at DESY (January 2009) and SLAC (May 2009). This document is an intermediate report to the International Committee for Future Accelerators (ICFA) of the reflections of this Study Group.

1

2012
Blueprint

arXiv:1205.4667


DPHEP-2012-001
May 2012

Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics

www.dphep.org

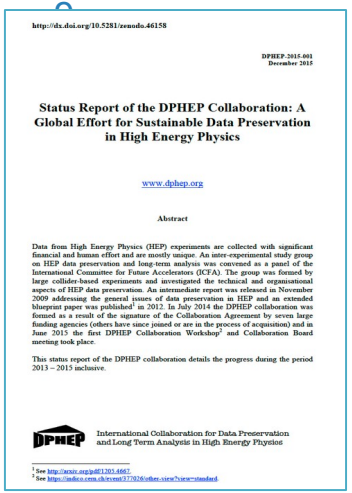
Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP. This paper includes and extends the intermediate report. It provides an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organisation in charge of the data management and policies in high-energy physics.

DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

2015
Collaboration MoU

arXiv: 1512.0201


http://dx.doi.org/10.5281/zenodo.46158

DPHEP-2015-001
December 2015

Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics

www.dphep.org

Abstract

Data from High Energy Physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP and an extended blueprint paper was published[1] in 2012. In July 2014 the DPHEP collaboration was formed as a result of the signature of the Collaboration Agreement by seven large funding agencies (others have since joined or are in the process of acquisition) and in June 2015 the first DPHEP Collaboration Workshop[2] and Collaboration Board meeting took place.

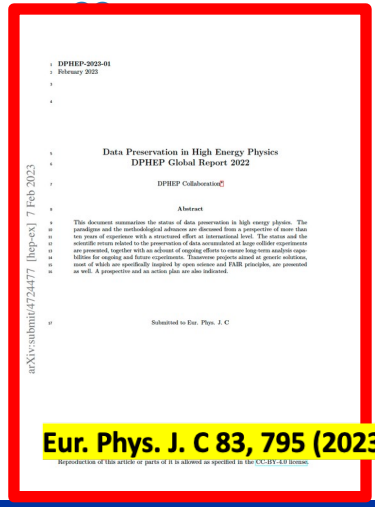
This status report of the DPHEP collaboration details the progress during the period 2013 – 2015 inclusive.

DPHEP International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

[1] See http://arxiv.org/pdf/1205.4667
[2] See https://indico.cern.ch/event/377026/other-view?view=standard

2023
Decade report

arXiv: 2302.035


arXiv:submit/4724477 [hep-ex] 7 Feb 2023

DPHEP-2023-01
February 2023

Data Preservation in High Energy Physics
DPHEP Global Report 2022

DPHEP Collaboration[*]

Abstract

This document summarizes the status of data preservation in high energy physics. The paradigms and the methodological advances are discussed from a perspective of more than ten years of experience with a structured effort at international level. The status and the scientific return related to the preservation of data accumulated at large-collider experiments are presented, together with an account of ongoing efforts to ensure long-term analysis capabilities for ongoing and future experiments. Transverse projects aimed at generic solutions, most of which are specifically inspired by open science and FAIR principles, are presented as well. A prospective and an action plan are also indicated.

Submitted to Eur. Phys. J. C

Eur. Phys. J. C 83, 795 (2023)

Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

2025
EPPS 2026 submission

2503.23619 [hep-ex]

arXiv:2503.23619v1 [hep-ex] 30 Mar 2025

DPHEP-2025-01

Data Preservation in High Energy Physics

DPHEP Collaboration

April 1, 2025

Abstract

Data preservation significantly increases the scientific output of high-energy physics experiments during and after data acquisition. For new and ongoing experiments, the careful consideration of long-term data preservation in the experimental design contributes to improving computational efficiency and strengthening the scientific activity in HEP through Open Science methodologies. This contribution is based on 15 years of experience of the DPHEP collaboration in the field of data preservation and focuses on aspects relevant for the strategic programming of particle physics in Europe: the preparation of future programs using data sets preserved from previous similar experiments (e.g. HERA for EIC), and the use of LHC data long after the end of the data taking. The lessons learned from past collider experiments and recent developments open the way to a number of recommendations for the full exploitation of the investments made in large HEP experiments.[1]

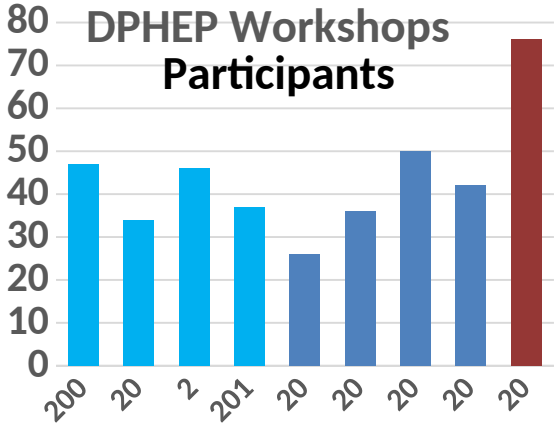
1 Introduction and context

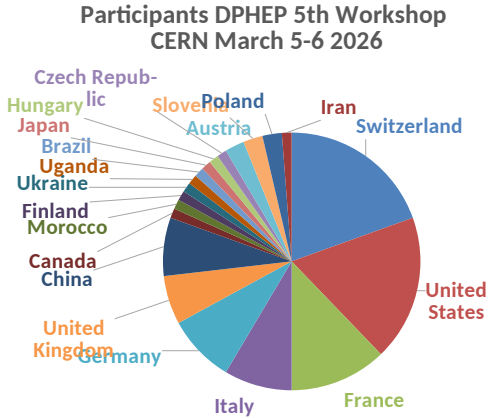
The need for data preservation demands long-term collaboration to support the continuing analysis program of complex data produced by large experiments. This was highlighted at the conclusion of operation of several large collider experiments at the beginning of this century, in the years before the LHC began. A study group was convened and produced in 2009 a report [1] that proposed a coordinated initiative in HEP, further detailed in a 2012 blueprint [2]. The document outlined initial data preservation projects and solidified the collaborative framework.

[1] This document has been submitted to the European Strategy for Particle Physics 2026.

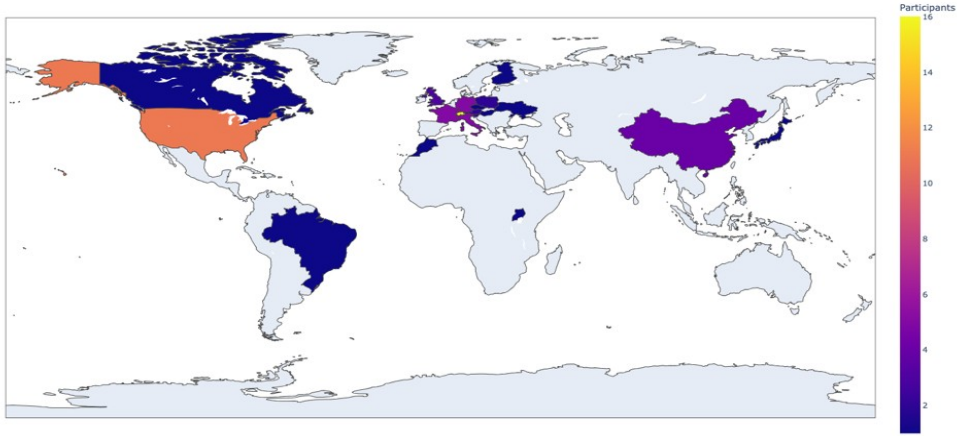
5th DPHEP Workshop 2026

- High participation increasing interest
  - 76 registrants, 26 abstracts
- International coverage
- Representation of ”known” experiments and sites



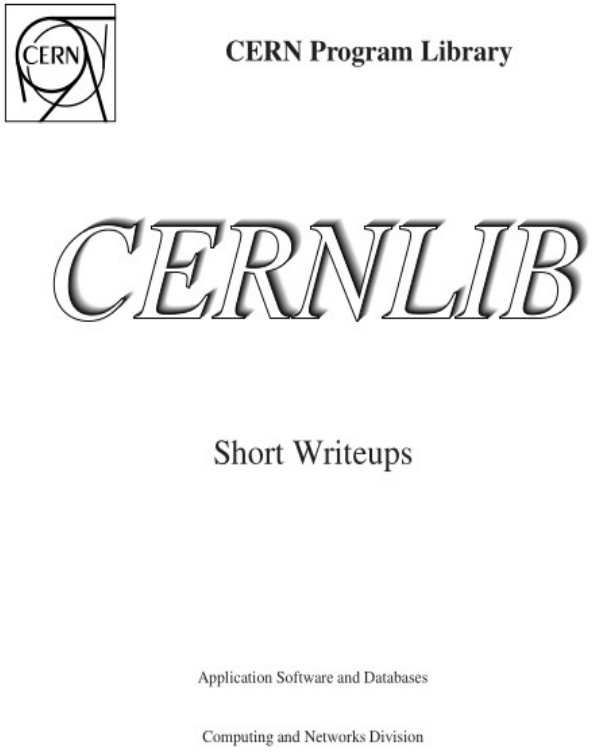


Participants by Country



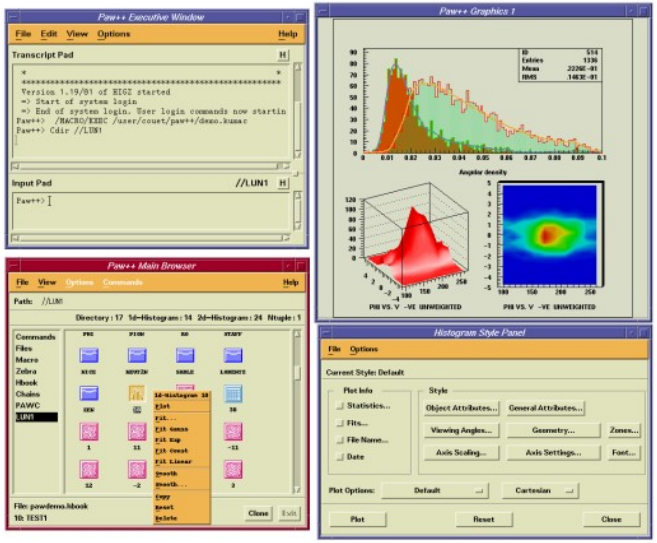
Past experiment tooling: CERNLIB and PAW

PAW = Physics Analysis Workstation



- Collection of tools
  - Maths, graphics, tooling
  - ZEBRA
  - …
- Used by numerous experiments at the time, e.g. LEP experiments
- Unsupported since 2006
- Community maintained version available since 2022
  - Works on 64bit, albeit with some limitations (*)
  - Used in data preservation, in particular for LEP experiments
- Features:
  - Allowing direct access to storage from the applications
  - Developed in and for 32bit
  - Developed on and ported to VAX and various UNIX flavors
    - DEC, ALPHA, HPUX
    - Later also Linux
  - **Optional interface with CERN storage systems via libshift**

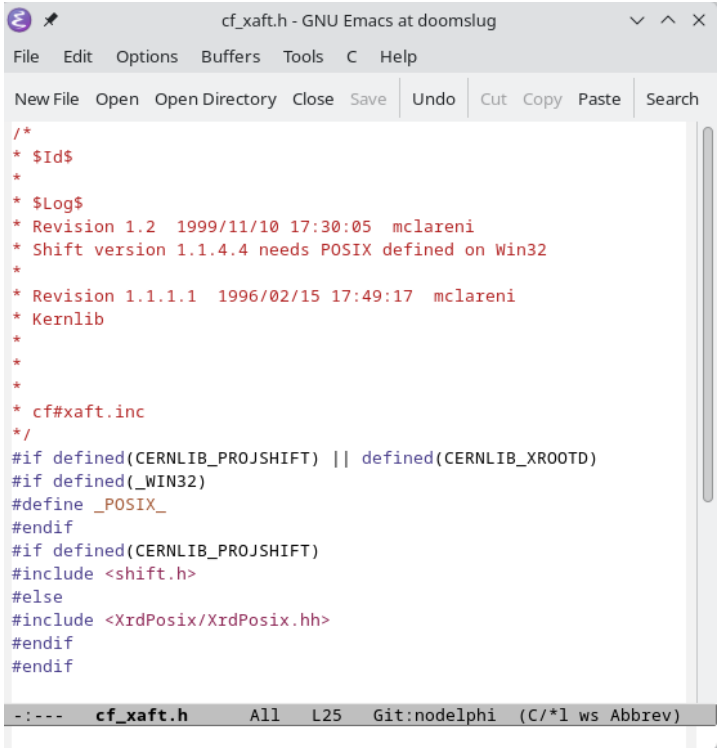
(*) See: Andrii Verbytskyi and Ulrich Schwickerath 2026 J. Phys.: Conf. Ser. 3206 012009



- Physics analysis workstation
  - PAW and PAW++
  - **Included in CERNLIB**
- Features:
  - Build-in Fortran interpreter
  - Ntuple processing (hbook)
  - Graphical display and printing
  - Predecessor of ROOT
- Still used by some people
  - Because of its simplicity ?
- Nice test case for data access

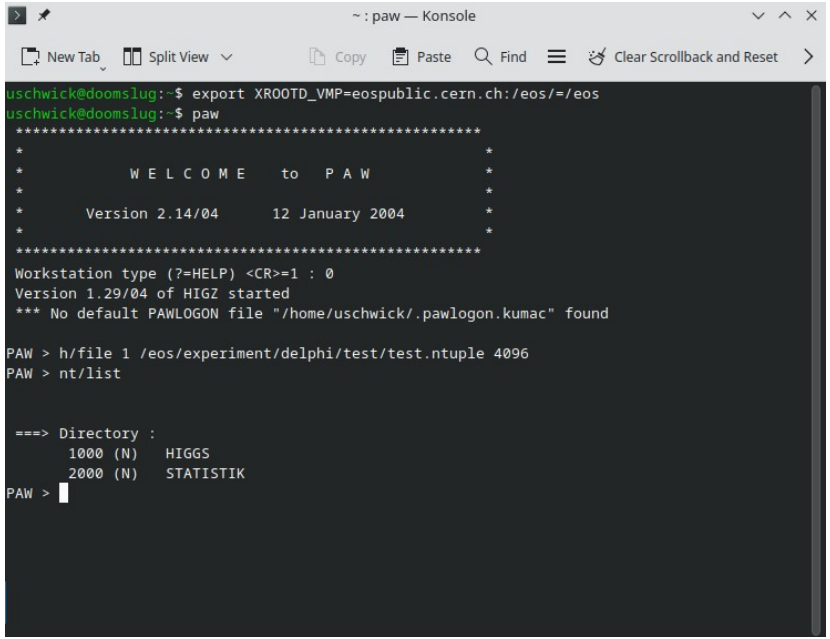
Interfacing CERNLIB with XrootD

- CERNLIB libshift interface
  - Implements support for RFIO via its POSIX interface
  - Intercepts system calls like open, read, write, close
- Very much the same as the XrootD POSIX interface!



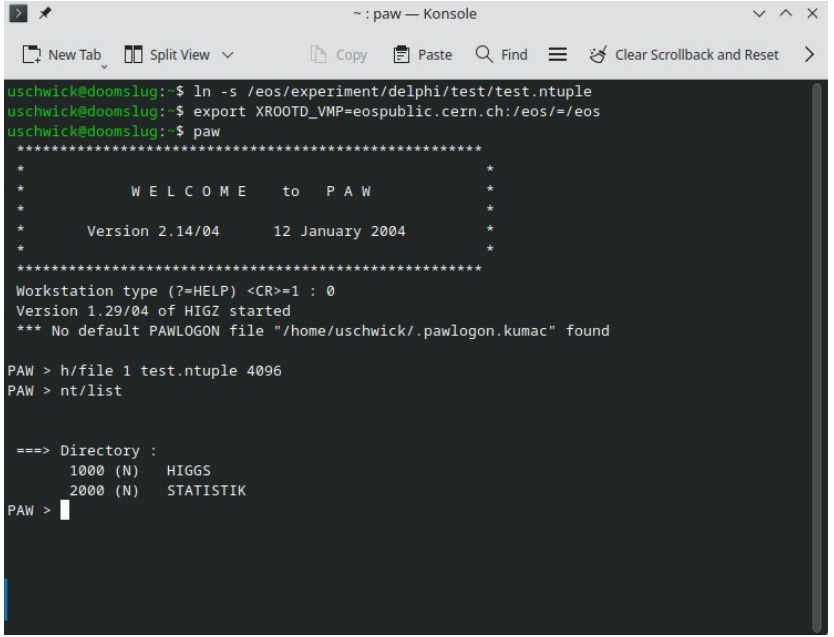
Interfacing CERNLIB with XrootD (5.x)

- Link it with XrdPosix
- Setup e.g. EOSPUBLIC
  - export XROOTD_VMP=eospublic.cern.ch:/eos/=/eos
- Tested with Kerberos authentication. Works.
- Tested with DELPHI stack. Works as well.
- Read a file from EOS over a local link



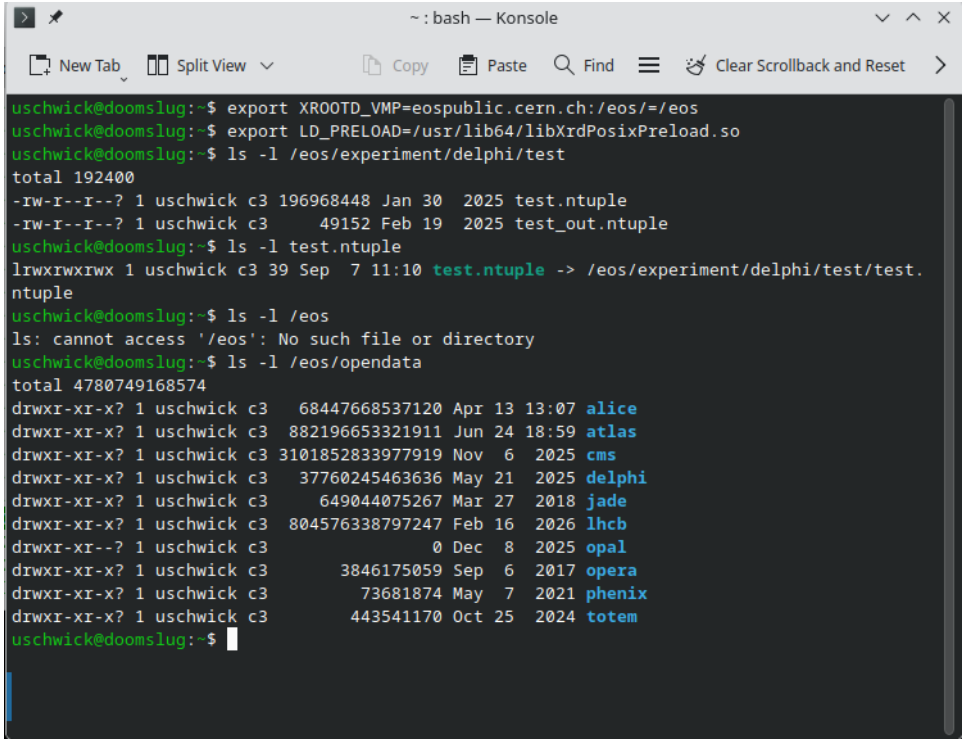
Improvements for XrootD (6.x)

- Following local links
  - Frequent use case in old experiment frameworks
- Implementation of missing statx systemcall
  - Used on Linux for e.g. ls
  - Enable use of LD_PRELOAD
  - Partly removes need to link with XrootD



Improvements for XrootD (6.x)

- Using LD_PRELOAD
  - No need to link with XrootD
    - export XROOTD_VMP=eospublic.cern.ch:/eos/=/eos
    - export LD_PRELOAD=/usr/lib64/libXrdPosixPreload.so
- Caveats:
  - Affects everything in the shell
  - May break other applications



Use cases and possible caveats

- Reading data directly over the network
  - Faster than mounting /eos
  - Eosclient is not always available for all Linux flavors
  - Can be used without having administrative privileges
  - Used already on Windows using WSL/WSL2 with DELPHI open data
- LD_PRELOAD of XrdPosixPreload
  - Eliminates the need to link with XrootD libraries
  - More resilient against XrootD version changes, e.g. the switch from XrootD 5 to 6 was not transparent and requires re-compilation of executables

Summary

- Preserving access to legacy data is important
  - Bridging the past to the future
- XrootD can be used with legacy software
  - Ensure continued data access from old frameworks
    - If these can still be run on modern machines
  - Demonstrated with CERNLIB as a prominent use case
  - Used by DELPHI and OPAL experiment frameworks
- The **XrootD POSIX interface** is very useful for this
  - Was not well maintained in the past
  - Numerous improvements in recent XrootD 6.1 and newer