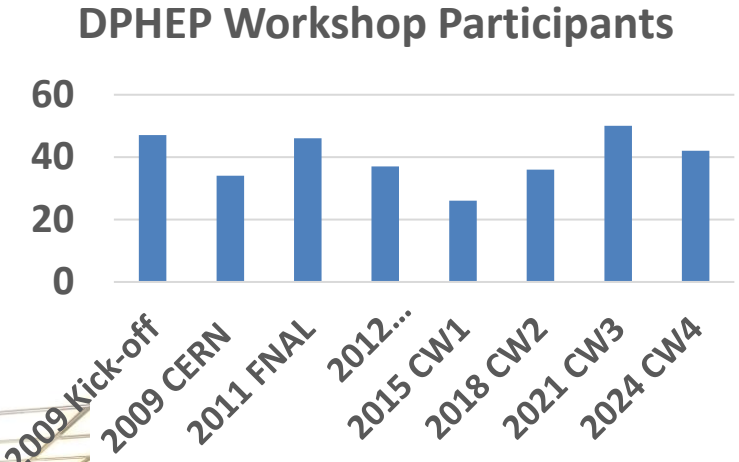
4th DPHEP Collaboration Workshop
Oct 2-3, 2024, CERN

C. Diaconu
DL Panel
Oct 8, 2024

https://indico.cern.ch/event/1432766/ report for DL panel



Workshop goals:

- • Review, update, discuss, plan
- • Progress report / survey
- • CHEP2024 Contribution
- • EPPSU Contribution/Statement
- • Organisation and future activity

Agenda

Wednesday Oct 2nd

- 14:25 **ALEPH**; Jacopo Fanini
- 14:45 **CERNLIB;** Andrii Verbytskyi, Ulrich Schwickerath
- 15:00 **DELPHI** ; Dietrich Liko, Dr Ulrich Schwickerath
- 15:15 **OPAL** ; Matthias Schroeder
- 15:30 **DELPHI and OPAL event displays**; M.Schroeder
- 15:45 → 16:00 Preserved Coffee
- 16:00 **ZEUS**; Achim Geiser
- 16:20 **H1** ; Speaker: Henry Klest
- 16:40 **JADE** Andrii Verbytskyi /Richard Hildebrandt
- 17:00 **PHENIX** ; Maxim Potekhin
- 17:20 **BaBar** ; Marcus Ebert

Thursday Oct 3rd

- 09:00 **KEK / Belle I & II** ; Takanori Hara
- 09:20 **BESIII** Gang Chen
- 09:40 **CERN Open Data portal** Pablo Saiz
- 10:00 **REANA** Marco Donadoni (CERN)
- 10:20 **CERN Analysis Preservation porta** P. Fokianos
- 10:45 → 11:00 Preserved Cofee
- 11:00 **CERN Open Data: Policy/implementation**; J. Boyd
- 11:20 **ALICE** ; David Dobrigkeit Chinellato
- **11:40 ATLAS**; Zach Marshall
- 12:00 **LHCb**; Dillon Fitzgerald
- 12:30 → 14:00 Preserved Lunch
- 14:00 **Preserving ANTARES legacy data** ; Jutta Schnabel
  14:20 PUNCH4NFDI ; Achim Geiser
- 14: **40 CMS** ; Julie Hogan
- 15:00 **ICFA Data Lifecycle Panel** ; Kati Lassila-Perini
- 15:25 **DPHEP Collaboration**

Highlights

- Significant progress has been made since the last workshop, both on data preservation and opening of data
- While CMS has pioneered the publication of open data, the other LHC experiments are rapidly catching up now: ATLAS, LHCb, ALICE
- Open data policies are increasingly applied also beyond LHC; useful synergy and *data opening calendar*
- **LEP Data is back!**

- In many contributions continued funding of DP was mentioned as an issue. An example is the BaBar experiment, whose software is running on outdated hardware but there is no funding to replace them
- Transfer across generations is visible
  - HERA ➔ EIC; LEP ➔ FCC
- Ongoing review of CERN OC3, which may offer some opportunities for DPHEP.
- The survey is encouraging, following a bottom up approach starting from the people involved in the practical work.
  - Document the output, echo to 10y report's open questions

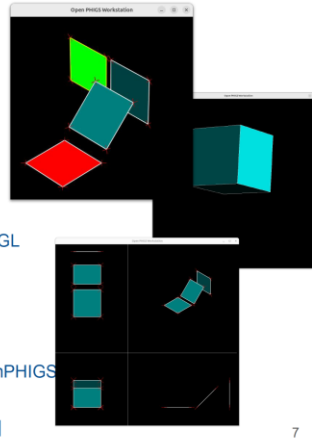
Highlights:LEP &co.

- LEP Data is back!
- Visual OPAL via rescued PHIGS alpha version
- ALEPH ➔ FCC
- CERNLIB

OpenPHIGS for data preservation

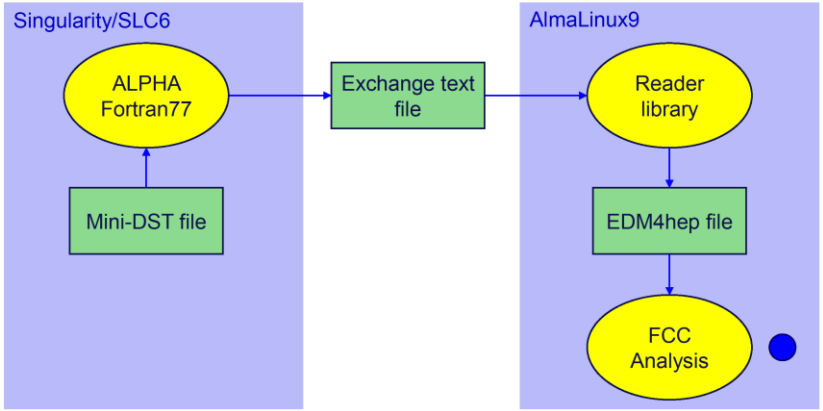
Caveats:

- Had to drop some functionality
  - Printing as PostScript is not possible
  - Replaced by TGA or PNG which is not the same resolution wise
- Still incomplete implementation
- Still bugs to be solved
  - E.g. filling only works for concave surfaces, as is the case in OpenGL
- Not fully compliant to the standard
- Still at pre-alpha level in terms of code maturity
- Still no documentation
  - Could import a subset of the man pages from the PEX based OpenPHIGS implementation
- Eventually not the same look and feel as with the original

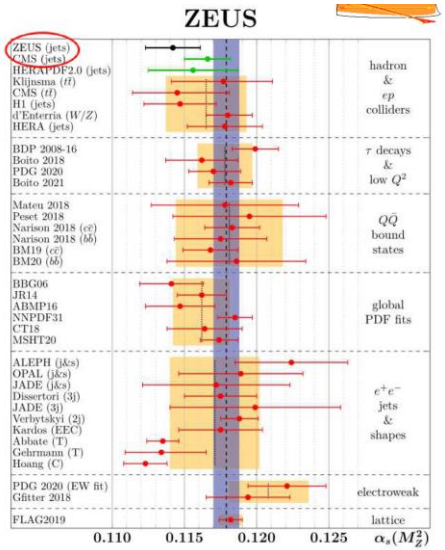


7

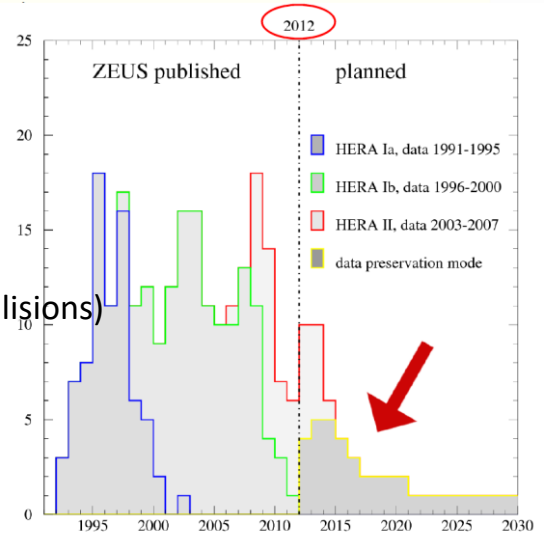
Workflow



HERA: olympic shape

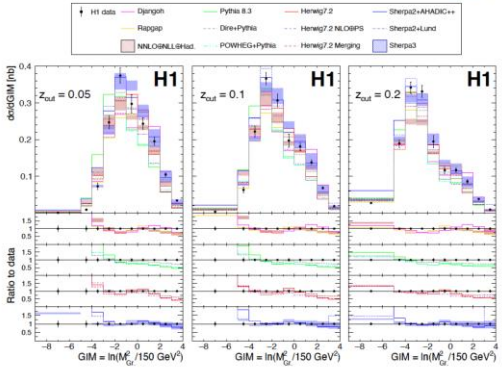


Two DP model, very similar succes
Attract new collaborators (17 years after the end of collisions)

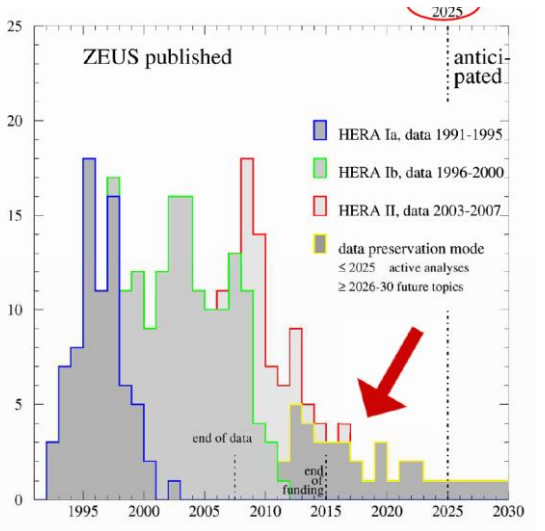


Recent physics results (2024)

- Recent H1 results: investigations of the hadronic final state
  - Jet substructure (2023)
  - Angular moments (2023)
  - Empty hemisphere events
  - 1-jettiness eventshape
  - Groomed eventshapes: [2403.10134], EPJ C84 (2024) 718



Groomed invariant mass and 1-jettiness are measured. First time grooming is tested in ep. Models have some difficulties to describe our data.



BaBar data: active, travelling, community support/help/rescue, fragile

Data Access

- Data available to analyses: ~1.5PB
  - no storage available at UVic for it
- GridKa agreed to host the BaBar data to be used by analyses
  - BaBar site since a long time
    - had already some data on site; anything missing was copied to GridKa
  - also the metadata db to find the data files needed in an analysis was already there
    - only update of content needed
  - BaBar environment configuration specifies XRootD and db access point for the data and db
- Framework at UVic needs to access data at GridKa via streaming...
  - works surprisingly well for normal event data
    - workflow: read event, process, read event, process,...
  - but conditions data is also read via streaming
    - large amount of data each job needs to read

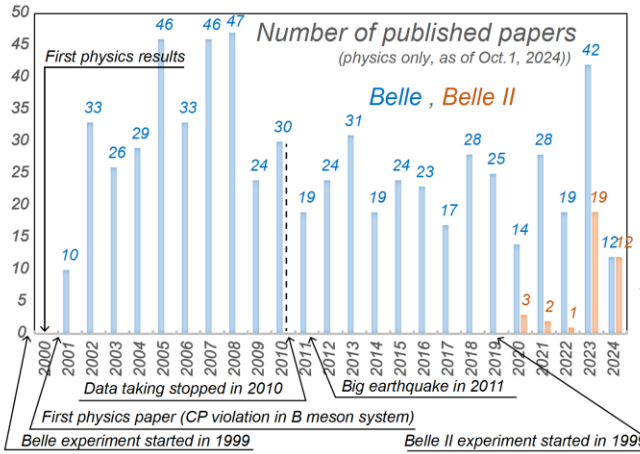
Hardware overview:

- XRootD proxy server: old machines
- XRootD redirector: VM on an old machine
- login machine: VM on an old machine
- BaBar interactive VM: VM on an old machine
- NIS server: VM on an old machine
- web server: on VM on an old machine
- babar wiki: VM on an old machine
- babar Hypernews: VM on an old machine
- NFS server: one new server, multiple old machines

Redundancy/Reliability:

- protect against disk failure
- protect against server failure

Active physics analysis with Belle data



Not in the left plot (Belle)
Accepted : 2
Submitted : 9
To be submitted : 7
(and many on-going analyses)

Even after 14 years, **Belle's data** is still being analyzed **vigorously**.

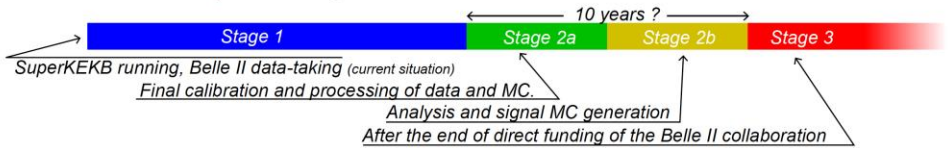
At the same time, **data analysis of Belle II** has become **more active**

DP / AP activities in Belle II

in June 2021, Belle II formed the Data Preservation Task Force

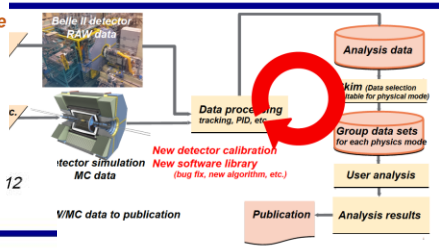
Charge

1. The expected impact of the data preservation plan on Belle II Physics publications,
2. The computing model required to enable the preservation plan, including raw data reprocessing and MC production, both in the post-SuperKEKB-running period and the post-Belle II lifetime,
3. The data that should be preserved,
4. The period of time for accessibility of the preserved data,
5. The analysis infrastructure that should be preserved,
6. The estimated cost and effort of Belle II data and analysis preservation, and
7. The outreach potential of open Belle II data



The Task Force presented Belle II with four priority recommendations for consideration
for detail, please check https://indico.belle2.org/event/7653/contributions/44071/

Difficulties on DP / AP



Extra Cost (budget, human)
Extra Computing resource

Extra Efforts
- Data archive with keeping accessibility
- Analysis preservation
- Common data format (HEP data is too complicated)

Fair evaluation in the collaboration ?
Career path ?

d of experiment dedicated direct funding,
will continue to manage them? (no-longer collaboration)

A sustainable "business" model that incorporates
long-term data / analysis preservation is necessary

v to make "data preservation" attractive for young researchers and/or researchers in other field e.g. informatics?)

Data preservation

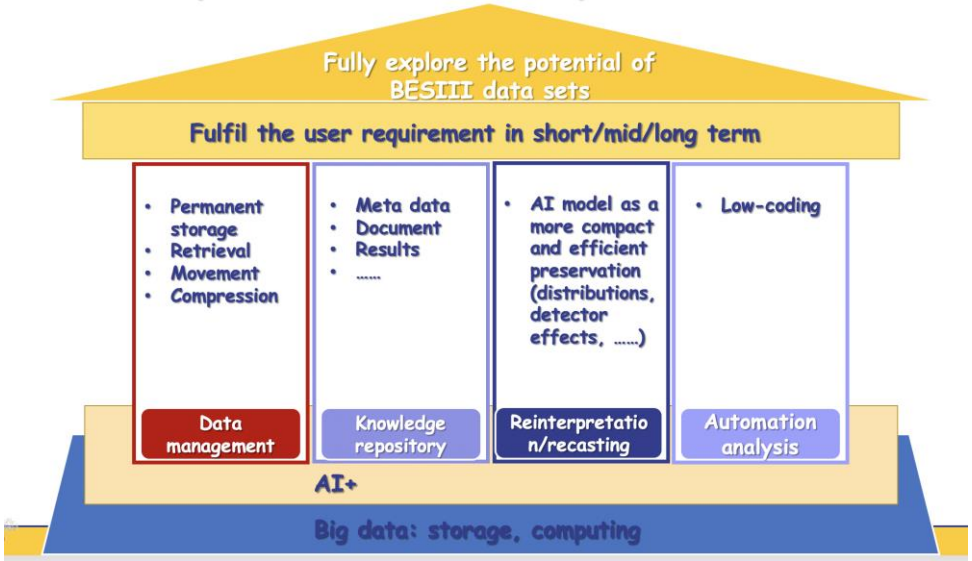
- BESIII adopts DPHEP Level-4 model
  - The full potential of data
    - RAW, DST (data, incl.MC),
    - metadata (calibration databases, ……),
    - software, documents
  - Adhere to the FAIR principles

DPHEP Collaboration: T. Basaglia, M. Bellis J. Blomer et al.: Data Preservation in High Energy Physics
Eur.Phys.J.C 83 (2023) 9, 795

Level	Model	Use Case
1	Provide additional information	Publication-related information search
2	Preserve the data in simplified form	Outreach, simple training analysis
3	Preserve the analysis-level software and data format	Full scientific analysis based on existing reconstruction
4	Preserve the reconstruction and simulation software and raw data	Full potential of the experimental data

BES III

AI-empowered data ecosystem for BESIII



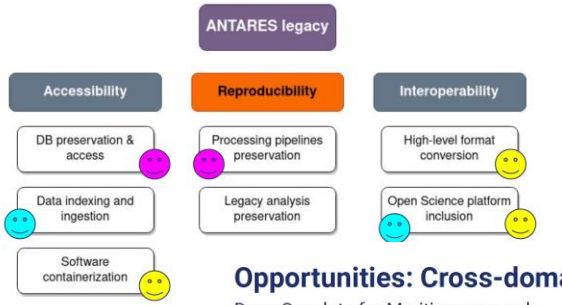
Antares : astro-neutrinos

Who is doing it

Funding and opportunities - first considerations

Funding options

- Partially integratable in KM3NeT infrastructure development (INFRADEV2)
- Included in ACME call (HORIZON-INFRA-2023-SERV-01), currently starting
  - *4.2.10. Access to neutrino data of ANTARES telescope.*
  - *4.3.2. ANTARES and KM3NET neutrino telescope data analysis services*
- Still looking for funding

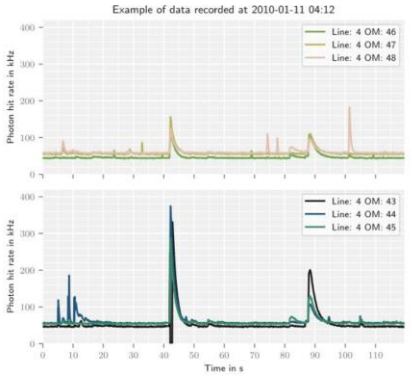


Opportunities: Cross-domain research

Deep-Sea data for Maritime research

Database preservation & access

- Database contains valuable information on environmental conditions in the Deep Sea and bioluminescence rate
- ORACLE database hard to preserve (supported versions, licensing ...)
- Considering containerization
- Aiming to provide interface for Maritime Science: DEEPSEA project @OSCARS, not funded



Studying bioluminescence flashes with the ANTARES deep-sea neutrino telescope. Limnol Oceanogr Methods, 21: 734-760. https://doi.org/10.1002/lom3.10578

Ongoing items

- Send abstracts to physics conferences (EPS2025, ICHEP)
- Structure and rhythm of meetings:
  - spring 2026 (before/with CHEP 2026)
- Survey on data preservation (Kati)
  - Feedback of the community to be documented as a report

CHEP 2024 DPHEP Talk

- Accepted on the basis of the 2023 report
- Speaker:
  - G. Ganis (CERN)

Data Preservation in High Energy Physics: a 10 years perspective

22 Oct 2024, 15:00
18m
Room 6

Talk | Track 8 - Collaborati... | Parallel (Track 8)

Speaker

Cristinel Diaconu (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR))

Description

Data Preservation (DP) is a mandatory specification for any present and future experimental facility and it is a cost-effective way of doing fundamental research by exploiting unique data sets in the light of the ever increasing theoretical understanding. When properly taken into account, DP leads to a significant increase in the scientific output (10% typically) for a minimal investment overhead (0.1%). DP relies on and stimulates cutting-edge technology developments and is strongly linked to Open Science and FAIR data paradigms. A recently released report (Eur.Phys.J.C 83 (2023) 9, 795 | 2302.03583 [hep-ex]) summarizes the status of data preservation in high energy physics from a perspective of more than ten years of experience with a structured effort at international level (DPHEP).

Primary author

Cristinel Diaconu (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR))

Presentation materials

There are no materials yet.

Work ahead

- Do we update the report? (shorter, arxiv)
  - or indico enough (well documented this time, thanks Ulrich for insisting on abstracts)
- EPSSU doc. (see thereafter)
- CERN Courier: contact established
- Refine/Publish the DPHEP Survey

https://cerncourier.com/a/data-preservation-is-a-journey/

https://cerncourier.com/a/study-group-considers-how-to-preserve-data/

ESPPU: a potential document structure

(if we are convinced that it is a sensible input to the ESPPU questions)

- General remarks on DP
- Scientific production
  - Exploiting the past
    - More science for low costs
    - b factories show that the analysis activity last longer and is more productive than the data taking periods
    - dedicated DP and ODS projects increase the scientific output
- Data Preservation and Open science
- Data Preservation and Technology
  - Robust data analysis sytems
  - New technologies being initiated such as AI –BesIII
- Community and cultural aspects of preserving (costly) HEP data sets
- Data preservation for the future; responsible treatment of public investment, training for future generations
  - Preparing the future LEP ➔FCC , HERA➔ EIC, b-factories legacy, LHC long term analysis
    - (note that scaling on the present experience, there will be publications from HL LHC data beyond 2050 at least) ,