CERN

Data preservation at LEP

Ulrich Schwickerath

Data preservation at LEP

- **Chapter I**: Data preservation introduction
- **Chapter II**: Aspects of data preservation
- **Chapter III**: Data preservation at LEP
- **Chapter IV**: Summary and conclusions

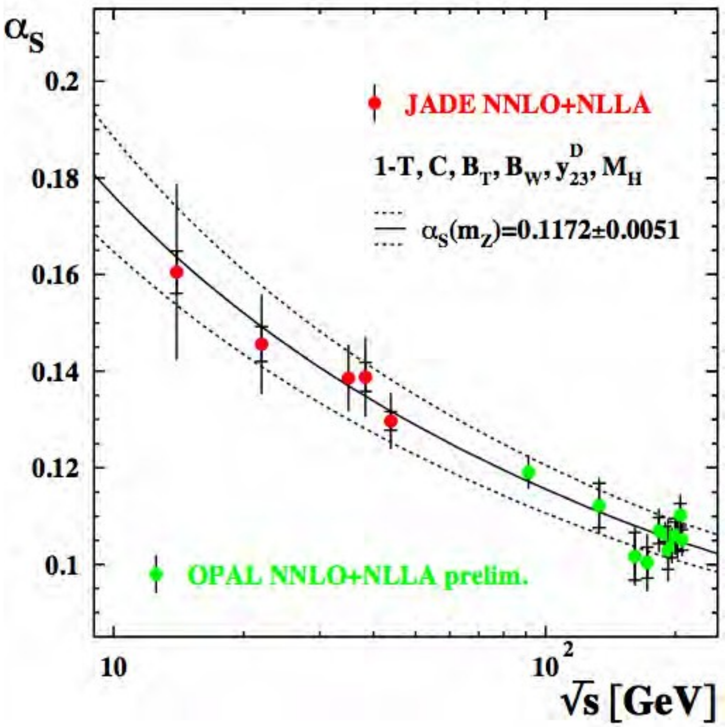
Introduction: the JADE story

The following slides are based on a

Talk by Siggi Bethke at KEK 2010

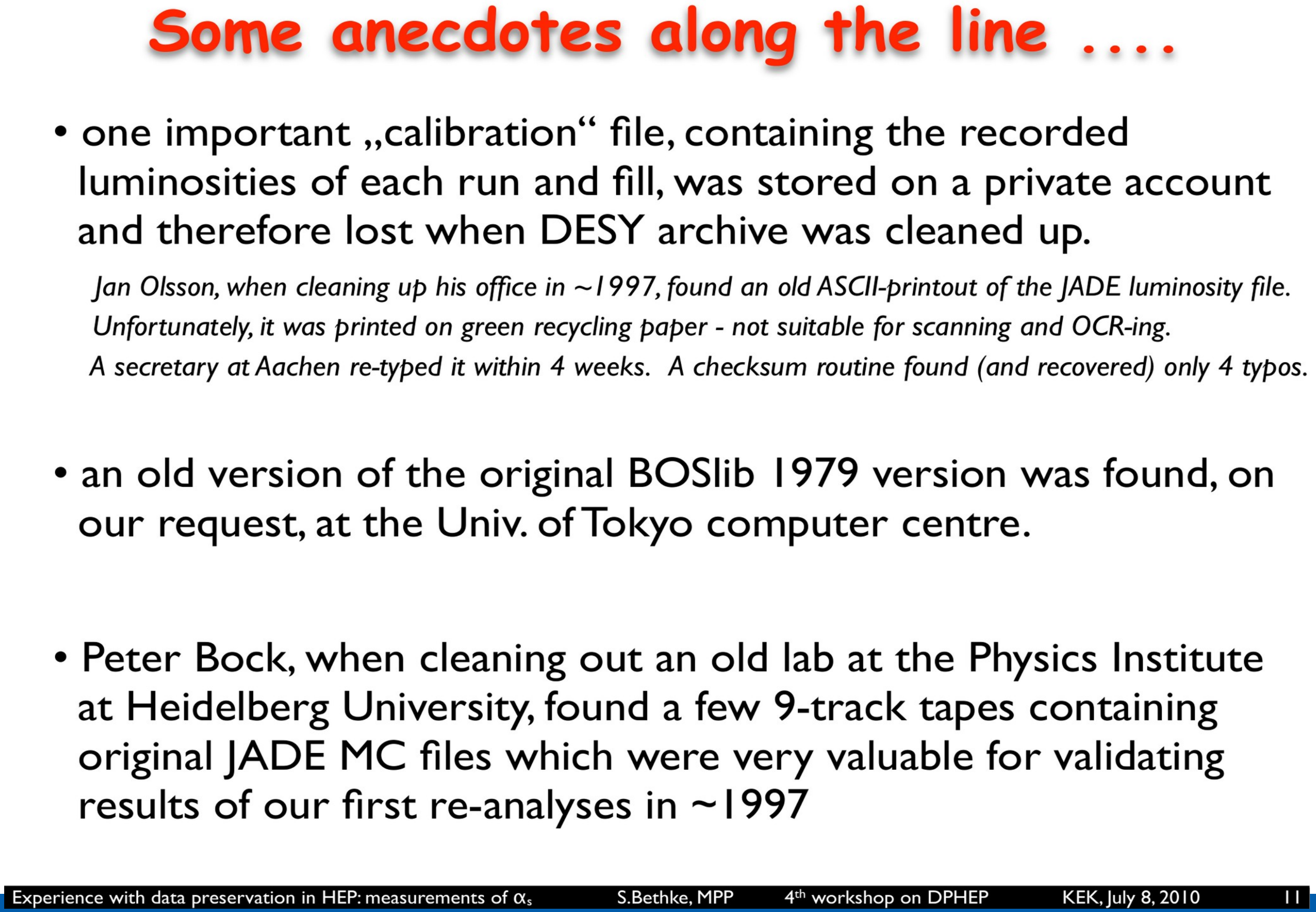
with his kind permission.

Re-analysis of JADE data



Courtesy: Siggi Bethke

- Re-vitalisation of **JADE software**
  - 1995 - 2003
  - Software partly rewritten
  - Full stack recovered
  - Simulation, reconstruction, event display
- Revitalisation of **JADE data**
- Result: **proof of Asymptotic Freedom**
- **See: Eur.Phys.J.C64:351-360,2009**
  - Siggi Bethke et. al.



Some anecdotes along the line ....

- one important „calibration“ file, containing the recorded luminosities of each run and fill, was stored on a private account and therefore lost when DESY archive was cleaned up.

  Jan Olsson, when cleaning up his office in ~1997, found an old ASCII-printout of the JADE luminosity file. Unfortunately, it was printed on green recycling paper - not suitable for scanning and OCR-ing. A secretary at Aachen re-typed it within 4 weeks. A checksum routine found (and recovered) only 4 typos.

- an old version of the original BOSlib 1979 version was found, on our request, at the Univ. of Tokyo computer centre.

- Peter Bock, when cleaning out an old lab at the Physics Institute at Heidelberg University, found a few 9-track tapes containing original JADE MC files which were very valuable for validating results of our first re-analyses in ~1997

Experience with data preservation in HEP: measurements of α_s S.Bethke, MPP 4th workshop on DPHEP KEK, July 8, 2010 11

Taken from a talk given by Siggi Bethke at KEK 2010,
With kind permission from S. Bethke

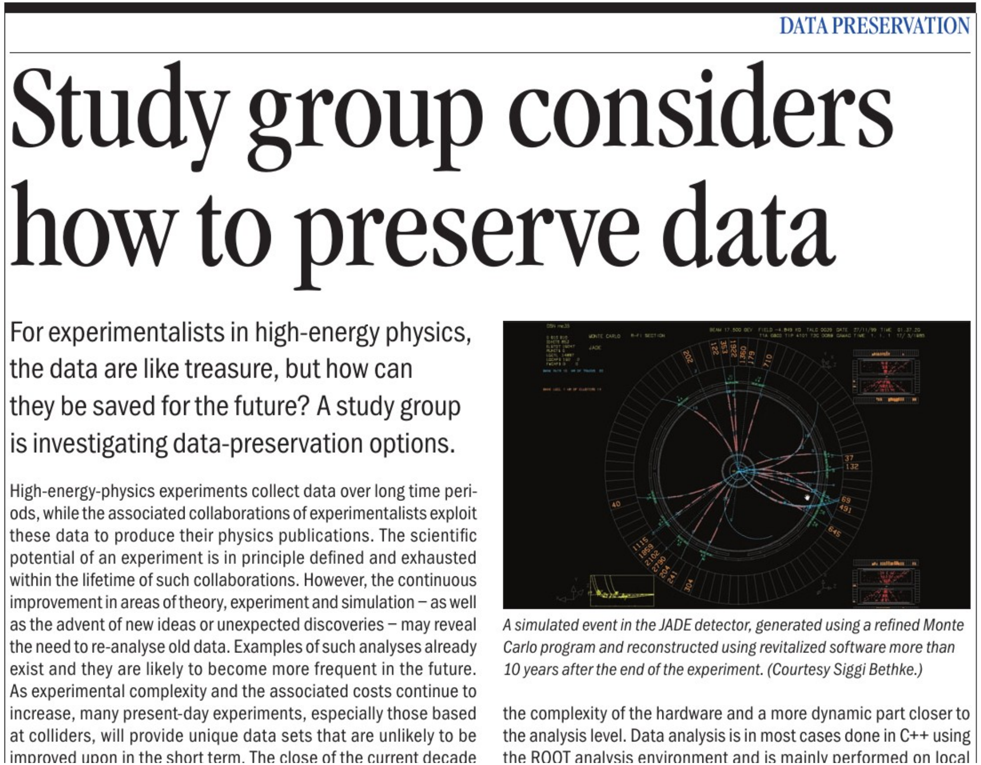
Lessons learned and consequences

Data preservation in high energy physics is important for

- Continued **improvements in theory, experiment and simulation** which may require a re-analysis of old data
- **New ideas** coming up
- Unexpected **discoveries**

DPHEP collaboration since 2009,

https://dphep.org



DATA PRESERVATION

Study group considers how to preserve data

For experimentalists in high-energy physics, the data are like treasure, but how can they be saved for the future? A study group is investigating data-preservation options.

High-energy-physics experiments collect data over long time periods, while the associated collaborations of experimentalists exploit these data to produce their physics publications. The scientific potential of an experiment is in principle defined and exhausted within the lifetime of such collaborations. However, the continuous improvement in areas of theory, experiment and simulation – as well as the advent of new ideas or unexpected discoveries – may reveal the need to re-analyse old data. Examples of such analyses already exist and they are likely to become more frequent in the future. As experimental complexity and the associated costs continue to increase, many present-day experiments, especially those based at colliders, will provide unique data sets that are unlikely to be improved upon in the short term. The close of the current decade

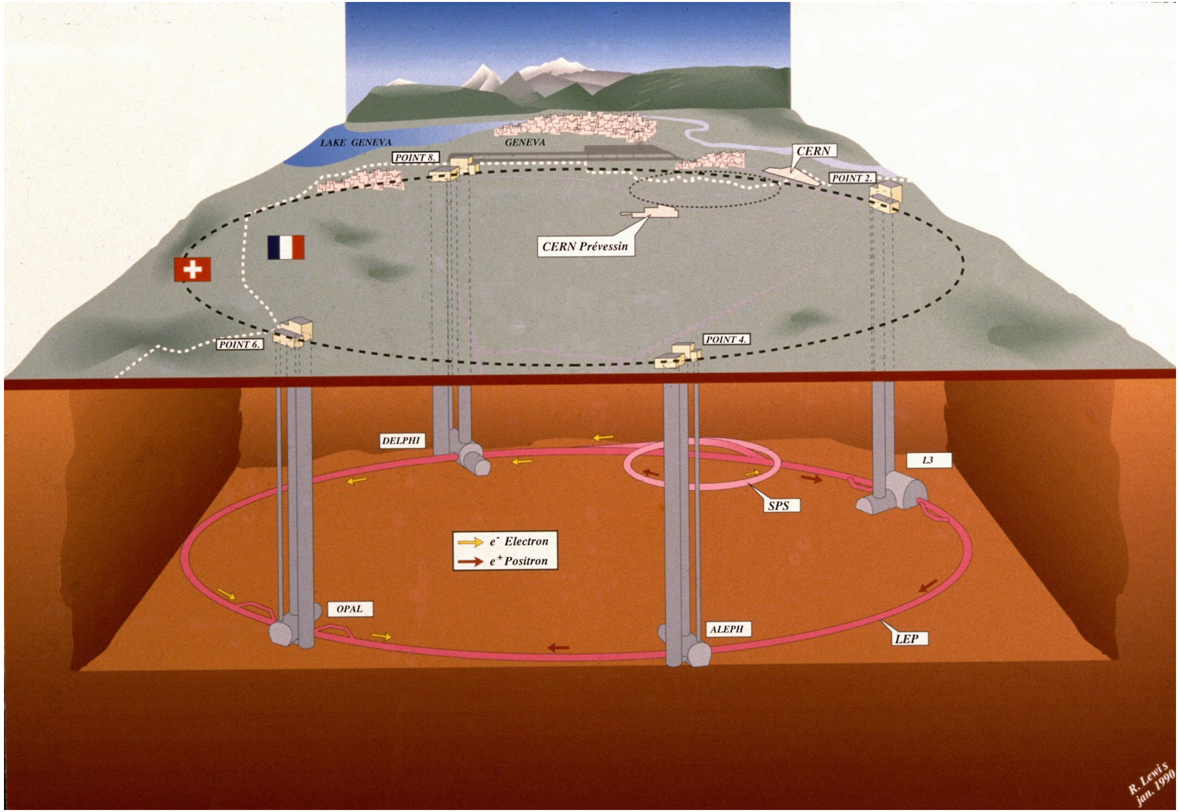
A simulated event in the JADE detector, generated using a refined Monte Carlo program and reconstructed using revitalized software more than 10 years after the end of the experiment. (Courtesy Siggi Bethke.)

the complexity of the hardware and a more dynamic part closer to the analysis level. Data analysis is in most cases done in C++ using the ROOT analysis environment and is mainly performed on local

CERNCOURIER 29 April 2009

What was LEP again ?

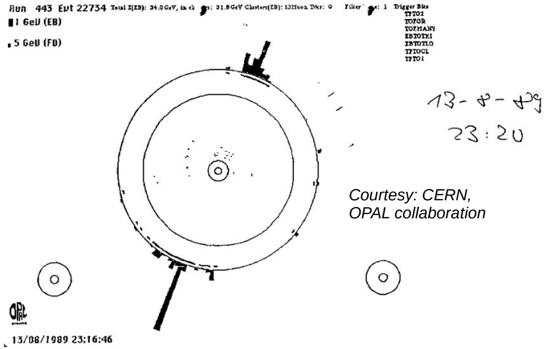
- e^+e^- collider at CERN
  - 1989-2000
  - 90-209 GeV
- 4 experiments:
  - ALEPH
  - DELPHI
  - L3
  - OPAL
- **Largest circular lepton collider so far**



Courtesy:CERN

LEP: some impressions

First Z0 event,
OPAL 1989


Run 443 Evt 22734 Total E(EB): 34.0GeV, in cl [illegible]: 31.8GeV Clusters(EB): 13Muon Trks: 0 Filter [illegible]: 1 Trigger Bits
TPTO2
TOFOR
TOFMANY
EBTOTHI
EBTOTLO
TPTOCL
TPTO1
1 GeV (EB)
5 GeV (FD)
13-8-89
23:20
OPAL
13/08/1989 23:16:46

Courtesy: CERN,
OPAL collaboration

Courtesy: CERN

LEP Tunnel
break through,
1986

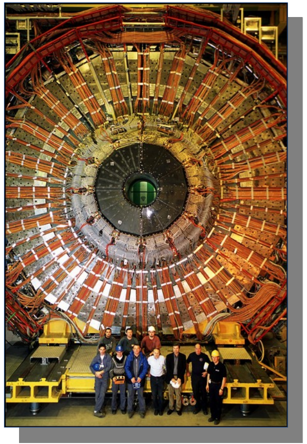
Courtesy: W. Liebig

LEP RF galleries

Courtesy: W. Liebig

LEP accelerator

Courtesy: CERN



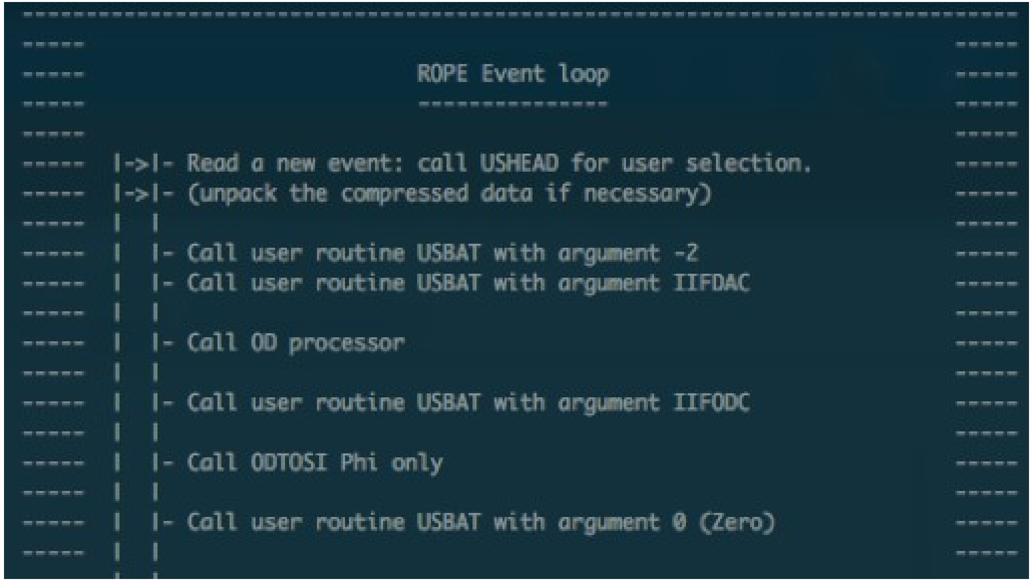
DELPHI end-cap before
dismantling

Challenges in data preservation in High Energy Physics (HEP)

- **Chapter I**: Data preservation introduction
- **Chapter II**: Aspects of data preservation
- **Chapter III**: Data preservation at LEP
- **Chapter IV**: Summary and conclusions

Aspects of data preservation

- Access policies
- Bit preservation
  - e.g. raw data and simulations
  - calibration databases, …
  - ntuples
- Software preservation
- Document Archival
  - Papers, notes, manuals, theses, ...
- Analysis preservation



```
--------------------------------------------------------------------------------
-----                                                                      -----
-----                          ROPE Event loop                             -----
-----                          ---------------                             -----
-----                                                                      -----
-----  |->|- Read a new event: call USHEAD for user selection.             -----
-----  |->|- (unpack the compressed data if necessary)                     -----
-----  |  |                                                                -----
-----  |  |- Call user routine USBAT with argument -2                      -----
-----  |  |- Call user routine USBAT with argument IIFDAC                  -----
-----  |  |                                                                -----
-----  |  |- Call OD processor                                             -----
-----  |  |                                                                -----
-----  |  |- Call user routine USBAT with argument IIFODC                  -----
-----  |  |                                                                -----
-----  |  |- Call ODTOSI Phi only                                          -----
-----  |  |                                                                -----
-----  |  |- Call user routine USBAT with argument 0 (Zero)                -----
-----  |  |                                                                -----
```

cernvm screenshot OPAL event loop, curtesy: Frank Berghaus

Challenges in data preservation in High Energy Physics (HEP)

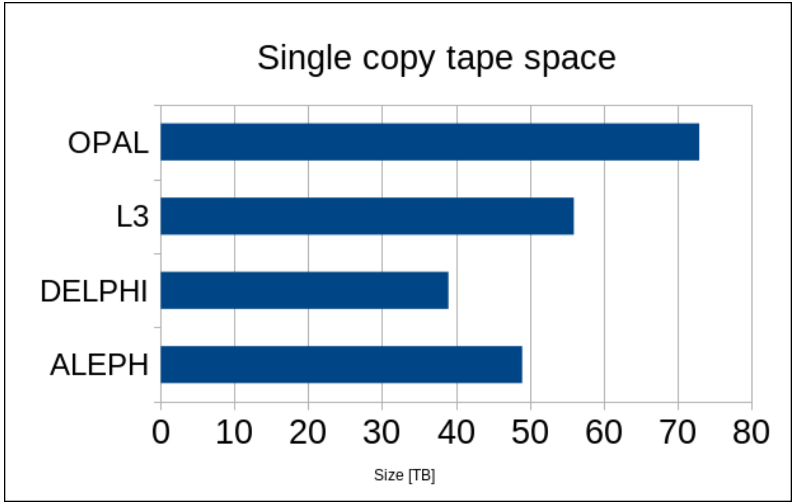
- **Chapter I**: Data preservation introduction
- **Chapter II**: Aspects of data preservation
- **Chapter III**: Data preservation at LEP
- **Chapter IV**: Summary and conclusions

LEP: data access policies

- Defines who can access the data
- Different for **each experiment**
  - Full data access is still handled relatively **restrictive**
  - Requires access to CERN and approval from the experiments
  - In some cases requires an experienced senior person from the experiment to assist in the analysis
- **Possible to access the data** for external people
- LHC: common data access policy document
  - https://cds.cern.ch/record/2745133/files/CERN-OPEN-2020-013.pdf

Bit preservation in general

- Long term storage of data, including
  - Raw data
  - Reconstructed data
  - Simulations
  - Databases, …
- Data format and representation
  - Typically **compressed binary data**
  - E.g. BOS, ZEBRA, ROOT
  - Human readable form would take too much space
- Technically considered to be a **solved problem**



Tape storage in use for LEP data

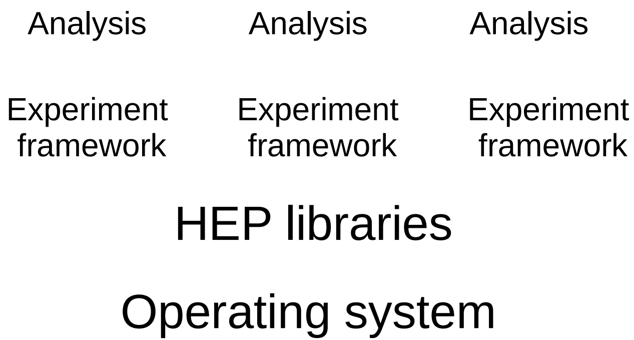
LEP: bit preservation

- LEP data on CERN Tape Archive (CTA) at CERN
  - **2 copies on tape**
  - Same for all 4 experiments
  - Raw data, detector data bases and simulations
- Additional copy on EOS at CERN
  - **Direct file based access**, e.g. from batch or interactive services at CERN
  - Full data set access usually protected
  - Some simplified data sets available on EOS, 4 vectors (e.g. ALEPH)
- **External copies** exist in some cases
  - E.g. DELPHI in Santander (Spain), for Opal in Munich
- **Open data initiative** for educational purpose,
  - Initiatives in ALEPH and DELPHI

Software preservation

The bare data is useless without the software to read it.

- Shared software between the experiments
  - HEP libraries, e.g. CERNLIB
  - Ongoing community effort to revive CERNLIB
- Experiment frameworks



Analysis Analysis Analysis

Experiment framework Experiment framework Experiment framework

HEP libraries

Operating system

Software preservation: options

Keep **software sources** alive

- Requires access to the full sources
- Keep adapting source code to changing computing environment
  - In version control system
  - As software CD with build scripts
- **Frequently rebuild and validate** the results
- Most flexible but also most **labor intensive** approach
- Avoids situations as for the Jade case

Note:

- Manual work in the past
- Today we have CI/CD which can help here (but they need to be setup first)

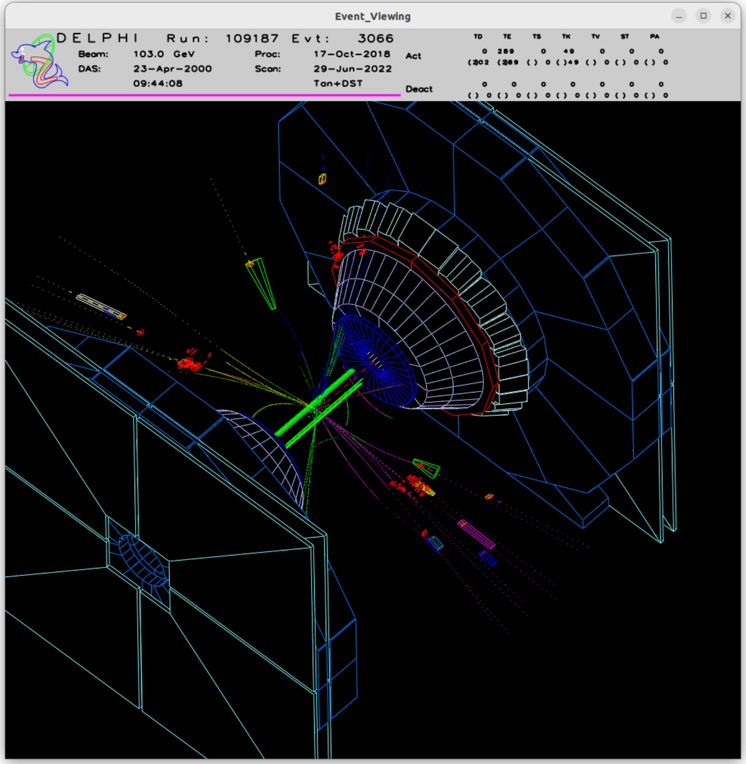
Software preservation: options

- Encapsulation and emulation
  - Install binaries on **physical hardware** and archive that
    - Likely to break physically over the years
    - Does not scale
  - Install binaries in **VM** or in a **container**
    - Or keep only the environment on the VM/container
    - Install the binaries on shared file system like AFS or cvmfs
    - No guarantee that the images will still work after many years, as technology changes
  - External dependencies and interfaces, e.g. with storage systems can be an issue
    - E.g. changing access protocols (like rfio which was used largely at LEP)
    - Generally, avoiding site dependencies is a good idea

LEP: software preservation

Experiment frameworks key components:

- **Analysis** frame work
  - User facing
  - Loop over events (data or simulated)
  - Run specific modules on them and
  - Execute user defined analysis code
- **Simulation**
  - Event generation
  - Simulation of the detector response
- **Reconstruction**
  - Reconstruct event based on detector response
  - Raw data or simulated raw data
- **Visualisation**
  - Event display, for interactive inspection of events
  - Can be useful for educational purpose as well



```
DELPHI   Run: 109187  Evt:   3066                     TD   TE   TS   TK   TV   ST   PA
   Beam:   103.0 GeV     Proc:   17-Oct-2018   Act     0  289    0   49    0    0    0
   DAS:    23-Apr-2000   Scan:   29-Jun-2022        (2)02 (2)89 ( ) 0 ( )49 ( ) 0 ( ) 0 ( ) 0
           09:44:08              Tan+DST
                                               Deact   0    0    0    0    0    0    0
                                                    ( ) 0 ( ) 0 ( ) 0 ( ) 0 ( ) 0 ( ) 0 ( ) 0
```

Challenge: Environment changes after 2000

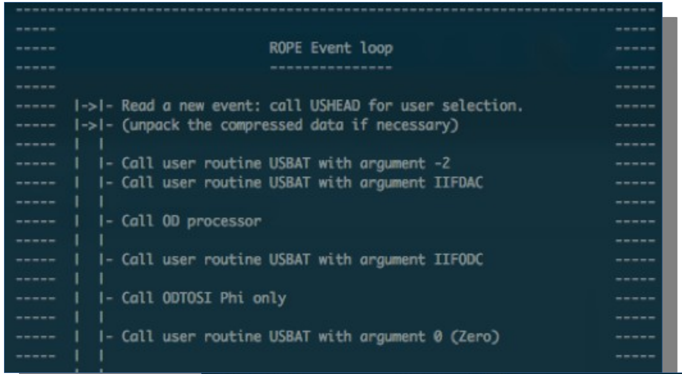
- Data storage and data access:
  - Fatmen (server part) stopped
  - RFIO support removed
  - Direct access to CASTOR removed
  - CASTOR replaced by CTA
- Computing
  - Other Unix flavors: stopped
  - CERNLIB support stopped
  - AFS areas removed
  - Dedicated web servers retired
  - Compiler changes (gcc)
  - g77 compiler abandoned
  - Commercial software libraries broken
  - License expired
  - Vanishing 32bit support

→

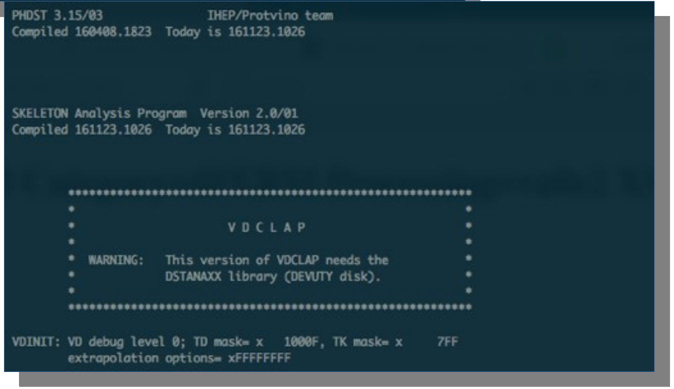
- New data storage services
  - EOS
  - CTA
  - CVMFS
- Computing
  - 64bit Linux
  - Multi-core CPUs
  - gfortran compiler
  - Different Linux flavors
- New services
  - Git and gitlab
  - Continous integration (CI)
  - Containers
  - EOS web
- Other
  - PHIGS prototype published as OpenSource on github

LEP: software preservation status

- Source code
  - Software CD (DELPHI)
  - Gitlab at CERN (DELPHI)
  - CVMFS (ALEPH, DELPHI, OPAL)
  - CERN VM (DELPHI, OPAL)
- Binaries
  - CVMFS (ALEPH, DELPHI, OPAL)
  - EOS (original binaries for DELPHI, ...)
- Computing environment
  - Originally on AFS
  - CERNVM mounting CVMFS (DELPHI, OPAL)
  - VM images (ALEPH, DELPHI, OPAL)
  - Docker images (DELPHI)
  - Archived PCs (in-official, DELPHI)



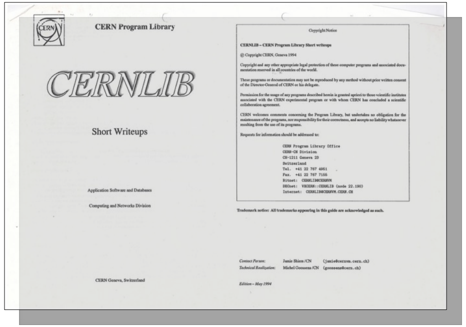
```
-------------------------------------------------------------------------------
-----                                                                     -----
-----                          ROPE Event loop                            -----
-----                          ---------------                            -----
-----                                                                     -----
-----   |->|- Read a new event: call USHEAD for user selection.           -----
-----   |->|- (unpack the compressed data if necessary)                   -----
-----   |  |                                                              -----
-----   |  |- Call user routine USBAT with argument -2                    -----
-----   |  |- Call user routine USBAT with argument IIFDAC                -----
-----   |  |                                                              -----
-----   |  |- Call OD processor                                           -----
-----   |  |                                                              -----
-----   |  |- Call user routine USBAT with argument IIFODC                -----
-----   |  |                                                              -----
-----   |  |- Call ODTOSI Phi only                                        -----
-----   |  |                                                              -----
-----   |  |- Call user routine USBAT with argument 0 (Zero)              -----
-----   |  |                                                              -----
```



```
PHDST 3.15/03             IHEP/Protvino team
Compiled 160408.1823  Today is 161123.1026



SKELETON Analysis Program  Version 2.0/01
Compiled 161123.1026  Today is 161123.1026


   *************************************************
   *                                               *
   *                  V D C L A P                  *
   *                                               *
   *  WARNING:   This version of VDCLAP needs the  *
   *             DSTANAXX library (DEVUTY disk).   *
   *                                               *
   *************************************************

VDINIT: VD debug level 0; TD mask= x   1000F, TK mask= x     7FF
        extrapolation options= xFFFFFFFF
```

LEP: the 64bit challenge

- **32bit support** is vanishing
  - At LEP time, **Linux based computers were running in 32bit mode only** (and usually single-core)
  - Nowadays, everything is 64bit (apart from some older gaming applications)
  - Example: **Motif 32bit libraries gone** from Ubuntu 20.04 and newer
  - Most LEP experiments rely on CERNLIB which is no longer supported for a while now
- **Community CERNLIB effort**
  - Recent effort to revive CERNLIB, in collaboration with MPG (Germany)
  - Good progress, aiming at a **first release in the next months**
  - Both 64bit and 32bit, enabling the experiments to migrate
  - Added support for cmake
  - Onboarding additional experiments started



CERN Program Library

CERNLIB

Short Writeups

Application Software and Databases

Computing and Networks Division

CERN Geneva, Switzerland

Copyright Notice

CERNLIB – CERN Program Library Short writeups

© Copyright CERN, Geneva 1994

Copyright and any other appropriate legal protection of these computer programs and associated documentation reserved in all countries of the world.

These programs or documentation may not be reproduced by any method without prior written consent of the Director-General of CERN or his delegate.

Permission for the usage of any programs described herein is granted apriori to those scientific institutes associated with the CERN experimental program or with whom CERN has concluded a scientific collaboration agreement.

CERN welcomes comments concerning the Program Library, but undertakes no obligation for the maintenance of the programs, nor responsibility for their correctness, and accepts no liability whatsoever resulting from the use of its programs.

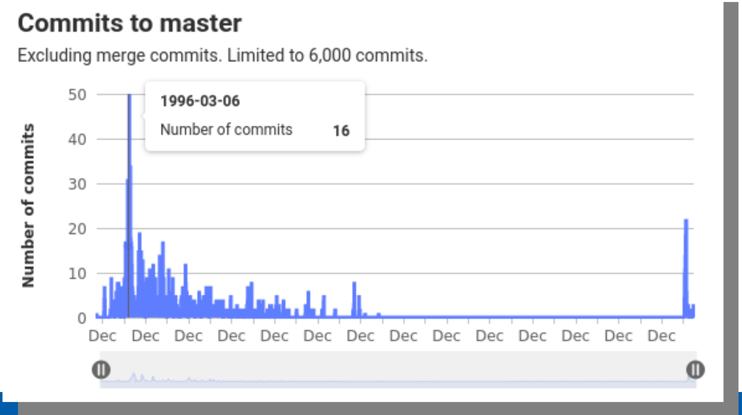
Requests for information should be addressed to:

```
CERN Program Library Office
CERN-CN Division
CH-1211 Geneva 23
Switzerland
Tel.  +41 22 767 4951
Fax.  +41 22 767 7155
Bitnet:  CERNLIB@CERNVM
DECnet:  VXCERN::CERNLIB (node 22.190)
Internet:  CERNLIB@CERNVM.CERN.CH
```

Trademark notice: All trademarks appearing in this guide are acknowledged as such.

Contact Person: Jamie Shiers /CN (jamie@cernvm.cern.ch)
Technical Realization: Michel Goossens /CN (goossens@cern.ch)

Edition – May 1994



Commits to master

Excluding merge commits. Limited to 6,000 commits.

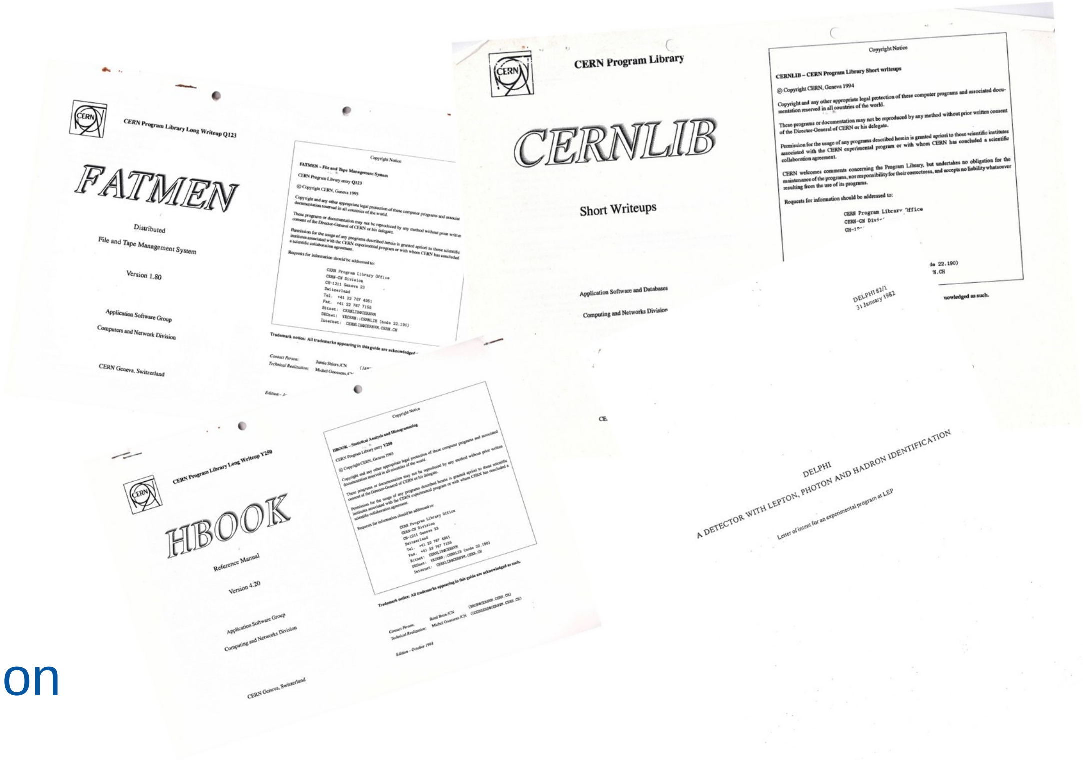
Knowledge preservation

Data and software are useless without the documentation on how to use them:

- Software without documentation is just another piece of data
- Knowledge preservation has to start early on in the process, before the experts leave, change, forget (or even die ...)

Knowledge preservation

- Manuals
  - Detector specifications
  - Software, interfaces, …
- Internal and technical notes
- Theses
- Conference contributions
- Publications
- Web pages and documentation



CERN Program Library Long Writeup Q123

FATMEN

Distributed

File and Tape Management System

Version 1.80

Application Software Group

Computers and Network Division

CERN Geneva, Switzerland

Copyright Notice

FATMEN – File and Tape Management System

CERN Program Library entry Q123

© Copyright CERN, Geneva 1993

Copyright and any other appropriate legal protection of these computer programs and associat documentation reserved in all countries of the world.

These programs or documentation may not be reproduced by any method without prior writte consent of the Director-General of CERN or his delegate.

Permission for the usage of any programs described herein is granted apriori to those scientific institutes associated with the CERN experimental program or with whom CERN has concluded a scientific collaboration agreement.

Requests for information should be addressed to:

```
CERN Program Library Office
CERN-CN Division
CH-1211 Geneva 23
Switzerland
Tel.  +41 22 767 4951
Fax.  +41 22 767 7155
Bitnet:  CERNLIB@CERNVM
DECnet:  VXCERN::CERNLIB (node 22.190)
Internet:  CERNLIB@CERNVM.CERN.CH
```

Trademark notice: All trademarks appearing in this guide are acknowledged

Contact Person: Jamie Shiers /CN (JAM

Technical Realization: Michel Goossens /C

Edition – J

CERN Program Library

CERNLIB

Short Writeups

Application Software and Databases

Computing and Networks Division

CE

Copyright Notice

CERNLIB – CERN Program Library Short writeups

© Copyright CERN, Geneva 1994

Copyright and any other appropriate legal protection of these computer programs and associated documentation reserved in all countries of the world.

These programs or documentation may not be reproduced by any method without prior written consent of the Director-General of CERN or his delegate.

Permission for the usage of any programs described herein is granted apriori to those scientific institutes associated with the CERN experimental program or with whom CERN has concluded a scientific collaboration agreement.

CERN welcomes comments concerning the Program Library, but undertakes no obligation for the maintenance of the programs, nor responsibility for their correctness, and accepts no liability whatsoever resulting from the use of its programs.

Requests for information should be addressed to:

```
CERN Program Library Office
CERN-CN Divi
CH-12
de 22.190)
N.CH
```

nowledged as such.

CERN Program Library Long Writeup Y250

HBOOK

Reference Manual

Version 4.20

Application Software Group

Computing and Networks Division

CERN Geneva, Switzerland

Copyright Notice

HBOOK – Statistical Analysis and Histogramming

CERN Program Library entry **Y250**

© Copyright CERN, Geneva 1993

Copyright and any other appropriate legal protection of these computer programs and associated documentation reserved in all countries of the world.

These programs or documentation may not be reproduced by any method without prior written consent of the Director-General of CERN or his delegate.

Permission for the usage of any programs described herein is granted apriori to those scientific institutes associated with the CERN experimental program or with whom CERN has concluded a scientific collaboration agreement.

Requests for information should be addressed to:

```
CERN Program Library Office
CERN-CN Division
CH-1211 Geneva 23
Switzerland
Tel.  +41 22 767 4951
Fax.  +41 22 767 7155
Bitnet:  CERNLIB@CERNVM
DECnet:  VXCERN::CERNLIB (node 22.190)
Internet:  CERNLIB@CERNVM.CERN.CH
```

Trademark notice: All trademarks appearing in this guide are acknowledged as such.

Contact Person: René Brun /CN (BRUN@CERNVM.CERN.CH)

Technical Realization: Michel Goossens /CN (GOOSSENS@CERNVM.CERN.CH)

Edition – October 1993

DELPHI 82/1
31 January 1982

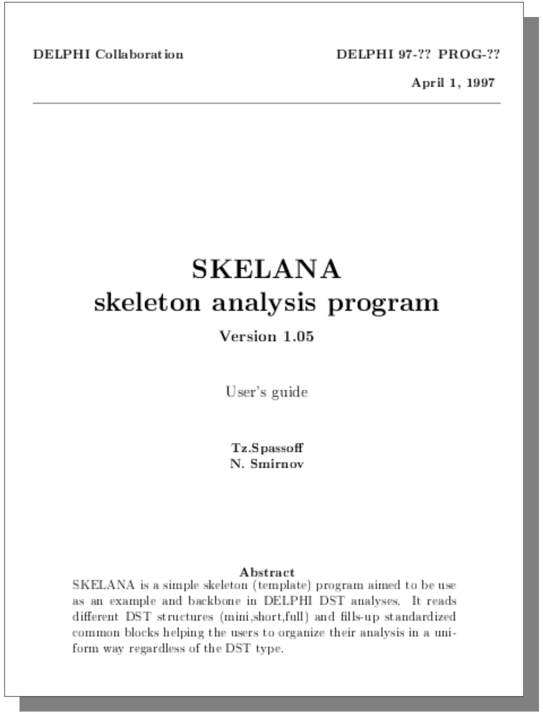
DELPHI

A DETECTOR WITH LEPTON, PHOTON AND HADRON IDENTIFICATION

Letter of intent for an experimental program at LEP

LEP: Knowledge preservation status

- Documentation - Manuals
  - CERN Document server (https://cds.cern.ch)
  - CVMFS (DELPHI, OPAL)
- Sample code, examples
  - CVMFS (ALEPH, DELPHI, OPAL)
  - Gitlab (DELPHI)
- Papers, internal notes, theses, photos…
  - CDS (https://cds.cern.ch)
  - Inspire (https://inspirehep.cern.ch)



DELPHI Collaboration

DELPHI 97-?? PROG-??

April 1, 1997

SKELANA

skeleton analysis program

Version 1.05

User's guide

Tz.Spassoff
N. Smirnov

Abstract

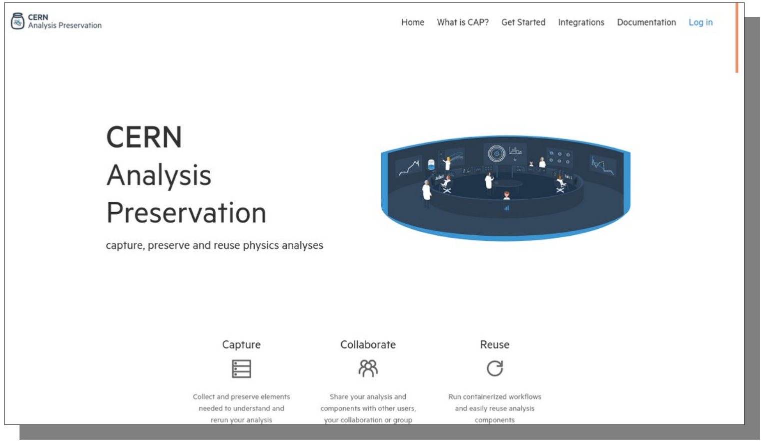
SKELANA is a simple skeleton (template) program aimed to be use as an example and backbone in DELPHI DST analyses. It reads different DST structures (mini,short,full) and fills-up standardized common blocks helping the users to organize their analysis in a uniform way regardless of the DST type.

LEP: Knowledge preservation status

- Early internal notes and technical documentation only on paper
  - No electronic copies available at the time
  - Physically archived e.g. in the CERN library
- Recovery attempt started in 2020 in DELPHI
  - About 2000 documents processed and checked
  - Scan, OCR process and archive electronically
  - Roughly 30000 pages in 1175 documents were recovered and imported into CDS at CERN
  - Mainly voluntary effort, in collaboration with the CERN library

Analysis preservation

- Archiving of Analysis
  - Requires **action during the run time** of the experiment already
  - A **common format** simplifies this a lot
  - Both software (sources) and building
  - Rapid re-run and reproduction of e.g. published results
  - **Re-use of existing and tested analyses** and processes
- Relevant for current experiments, e.g. LHC
  - Tools available now like reana, CAP, notebooks, ...
  - Too late for LEP ...



CERN Analysis Preservation

Home | What is CAP? | Get Started | Integrations | Documentation | Log in

CERN Analysis Preservation

capture, preserve and reuse physics analyses

Capture

Collect and preserve elements needed to understand and rerun your analysis

Collaborate

Share your analysis and components with other users, your collaboration or group

Reuse

Run containerized workflows and easily reuse analysis components

reana

Reproducible research data analysis platform

LEP: Analysis preservation

No general analysis preservation was done for LEP

- Some code and execs preserved, along with their output (ALEPH, DELPHI, OPAL)
  - Vanishing expertise how to run these
  - Not always obvious if the conserved code really matches to conserved output
  - May be useful for validation of re-compiled stacks
- **Missing documentation** and meta data on how to interpret these sets

Lessons to be learned from LEP

- General status
  - ALEPH, DELPHI and OPAL still in a fairly good shape, situation for L3 is a bit unclear
  - No strict analysis preservation has been done
- Avoid site dependencies as much as possible
- Avoid commercial and closed source software
  - DELPHI and OPAL used a commercial 3d visualisation tool called GPHIGS
  - New versions to adapt to new glibc and compiler versions cost money, if available at all ...
- Keeping the full software stack alive gives highest flexibility
- Keep and maintain a full software validation chain from the start
  - Implement automatic rebuilds and tests at all levels
  - CI can help here

Ongoing work

Developments around **software preservation**:

- **Community CERNLIB**, https://gitlab.cern.ch/dphep
  - Revival of CERNLIB (for Linux mainly)
  - Review and incorporation of plenty of upstream patches (Debian)
  - Use of gitlab CI to rebuild **32bit and 64bit versions** since SL4 to CentOS9
  - Ready for evaluation
  - First release hopefully later this year
- Removal of dependencies on commercial packages
  - Revival of the DELPHI event display
  - Maybe OPAL later

Ongoing work

Developments around **software and knowledge preservation**:

- Complete the move out of AFS
  - OPAL: ongoing, ALEPH, DELPHI: done
- Recover more software sources
  - Analysis code (OPAL, DELPHI)
  - Event display (OPAL)
- Migrate to newer OS versions
  - Support for 64bit and updated compilers
  - Requires validation
- Validation of recent libraries and community CERNLIB
- Recover paper copies, digitize and archive them
  - Ongoing/completed for DELPHI, not started for ALEPH and OPAL ?

Challenges in data preservation in High Energy Physics (HEP)

- **Chapter I**: Data preservation introduction
- **Chapter II**: Aspects of data preservation
- **Chapter III**: Data preservation at LEP
- **Chapter IV**: Summary and conclusions

Summary and conclusions

- There are many reasons why to preserve data in HEP
- DP needs to **start while the experiments are still alive** and taking data
  - At all levels
  - Early on at the design phase of the software
- Lessons can be learned from past attempts to recover data.
  - **DPHEP** collaboration for exchange of experience
  - Lots of **new tools** available nowadays to help preserving experimental results
  - **Efforts continue after the end of the experiment** to keep data access alive
  - Long term data preservation **is not for free** and needs funding
- Long term data preservation is **not restricted to HEP**.

Questions ?

www.cern.ch