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

Data Preservation (and more) at CERN

LEP (1989-2000), LHC(2009-40+)

CEOS WGISS-46

https://indico.cern.ch/event/763278/

Jamie.Shiers@cern.ch

30 years in 30 minutes!

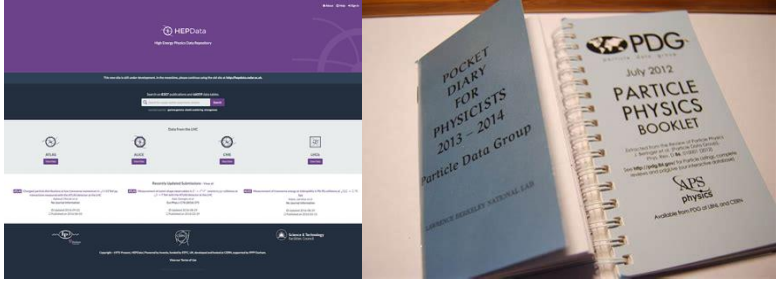
Suggested topics…

- Archive Infrastructure and technology
- Monitoring tools
- Archive volume, missions, coverage
- Archiving flows and processes
- Data format/packaging for long term archive
- Management of the relevant associated knowledge
- Volume and Trend
- AOB

Overview

1. Current state of Long-Term (multi-decade) Data Preservation at CERN
   - What data (“missions”) this covers
   - The main services involved including costs
2. How we got here (it was not (at) all planned)
   - Some key milestones & contributing factors
   - Major migrations & some dead-ends
3. Where we expect to go in the coming years

1. Current State



- About a dozen active experiments at CERN
  - Historically ~1000 since birth of CERN (1954)
- Most attention goes to the 4 main LHC experiments: ALICE, ATLAS, CMS, LHCb
- These together generate **50PB** (and growing) of data per data-taking year
- **Current "archive" around 300PB**
- Raw data is stored (tape) at CERN with a copy across ~10 "Tier1" centres

1. The Large Hadron Collider

- Originally proposed in late 1970s (78 / 79)
- “Real” data taking started in 2009/10
- Announcement of **Higgs Boson** in 2012
- Approximately 100 days of beam time per year with longer technical stops in the winter
- And occasional “long shutdowns” – LS2 starts imminently and will last until mid-2021
- Major machine (and detector) upgrades during LSx: after LS3: HL-LHC

➢ **These come with (big) increases in data rates**

1. Open Data at the LHC

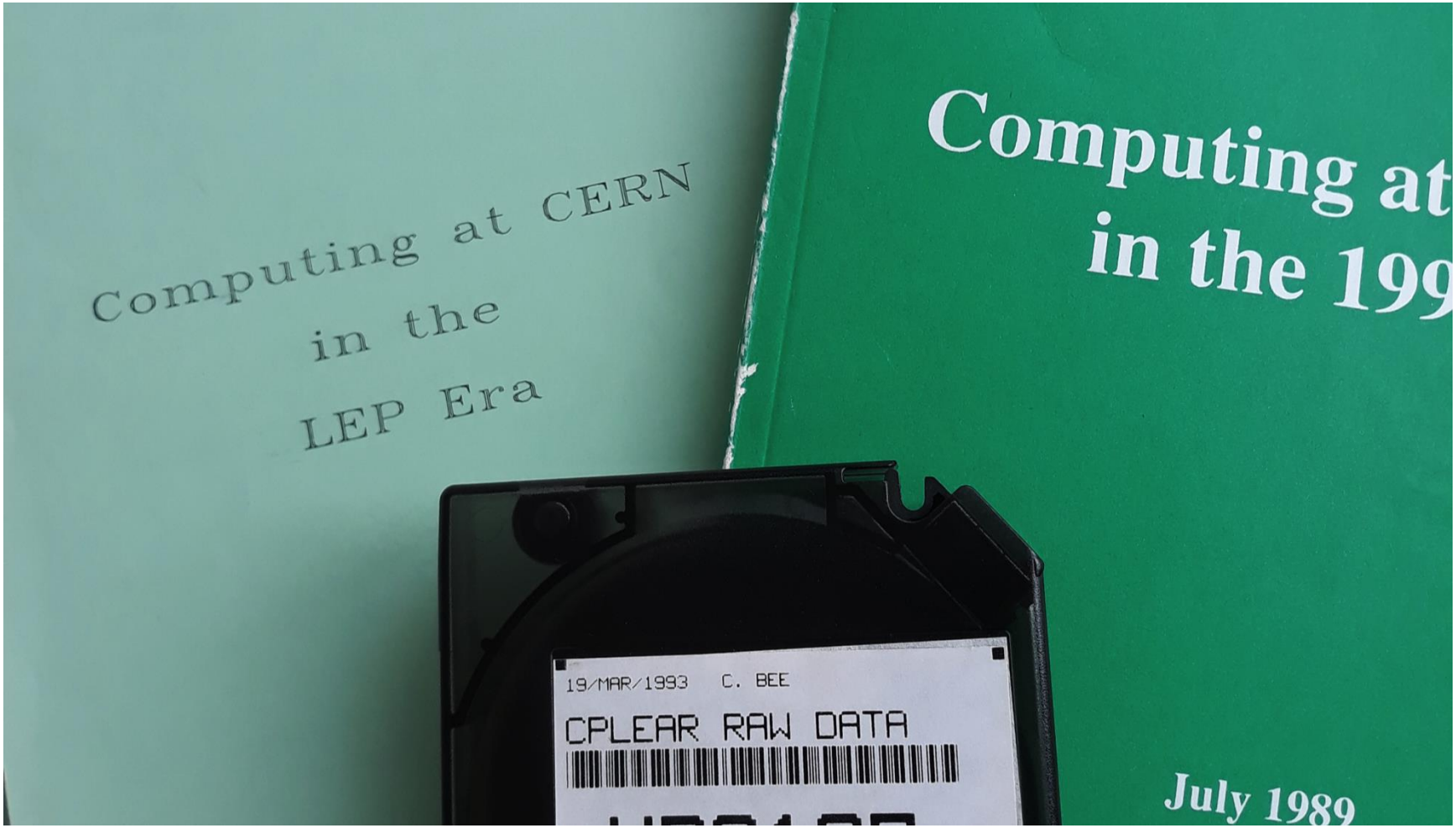
- For the first time at CERN / in HEP: the LHC experiments have Open Access policies
- **And have released more than 1PB of data**
- With the necessary software + VMs + documentation
- And it works – and is used!
- ➢ **Designated communities include "theorists", citizen scientists & high-schools**
- Big questions about funding for Open Data resources: but does it **ALL** need to be online **ALL** of the time?
- **Also – for the first time – preservation / curation of data part of an MoU signed by Funding Agencies**

1. Data Preservation in the LHC Era

- Key components described in the 2012/13 update of the **E**uropean **S**trategy for **P**article **P**hysics
- By 2016, these were matched by stable, production services (see iPRES paper)
- In 2017, task force looking at challenges of HL-LHC (2025+) and beyond concluded *inter alia* that ***"bit preservation with an acceptably low error rate can be considered a solved problem".***

➢ **Today, the focus has moved on (later) and another ESPP update is about to start…**

2. How did we get here?



2. The LEP Era

- Prior to LEP (1989-2000), experiments tended to be relatively short-lived
- LEP required a degree of planning that was previously unknown (but predictions wrong…)
- And it faced almost constant migrations (operating systems, computer hardware), as well as paradigms
  - Mainframe – clusters – farms – grids – clouds – ?
- **IMHO this need for regular migration is one reason why the data is still usable today**

- Another is the IBM 3480 cartridge… **and people…**

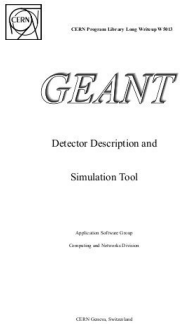
2. LEP Tape Management

- Experiments bought and managed their tapes: data was referred to by VSN+FSEQ
- **My contribution: introduction of a platform-independent naming scheme (file catalog) + the F and A in F.A.I.R. (maybe)**
- ***//catalogue/experiment/dir1/dir2/.../dirn/filename***
- This started the journey to “managed storage” where now all data is in robots
- Multiple copies of LEP data: 2 tape + 1 disk at CERN alone (extra copies at outside labs)

➢ **Migrating the data forward…**

2. LEP Software

- Software designed in an era of very heterogeneous computers & operating systems
- ***NEW*** Key decisions on machine independent data formats (e.g. IEEE floating point) & self-describing (to an extent) data
- Many migrations helped weed out bugs in code and ended up with 64 bit code for Unix / Linux
- Relatively(?) easy to port forward & validate
- **Try doing that with legacy mainframe code**
- ➢ **Porting the software forward…**



2. LEP Documentation

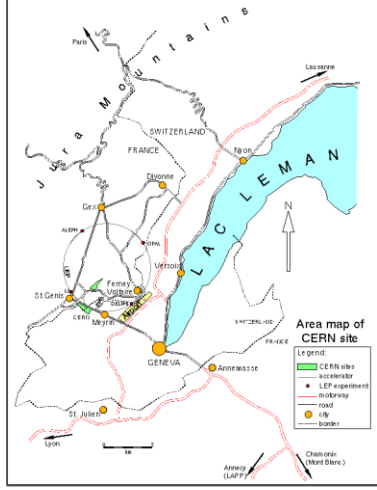
- Major effort in mid-1990s to re-format and bring up to date CERN Program Library documentation
  - LaTeX source, Postscript and HTML Output
- Re-done in ~2015 to produce consistent PDF, PDF/A and HTML – and capture much missing meta-data – and store in a Digital Library
  - DOIs – we did not go as far as ORCID IDs
- **Experts are needed to validate output!**

➢ **No automatic conversion of document formats**

2. PV2018 Proceedings

- I submitted an abstract, based on “Collaborative Long-Term Data Preservation: From Hundreds of PB to Tens of EB” to PV
- Somehow, the 4-page paper got “extracted” from the conference proceedings and fed into the CERN Document Server CDS
- It looks like it was **printed** & then **scanned** on the lowest resolution scanner left in the world
- The abstract was re-typed? OCR-ed?
- **DPHEP** was changed to **DPI-IEP**, even though the HEP acronym was defined in the same sentence

2. LEP Summary



- Started with mainframes, user-managed tapes and “unaffordable computing / storage”
- Profited from RISC and then PC revolutions
- Data preservation was not considered from the start – several unsuccessful attempts all doomed to failure for reasons including:
  - **Bit preservation “as a service” not available**
  - **Incomplete understanding of preservation**
  - **“Long-term” support & funding lacking**
- Despite this, the data from all 4 experiments is still available and that from 3 fully usable!

➢ **“Last gasp” data preservation – at the end, or even after, the end of data taking**

2. LEP Era: Tape Archive

- Ignore problem: we'd like to but….

~300K tapes were 'archived'…

.. ~150K were manually mounted…

..and then copied to Redwoods….

2. Preparing for the LHC

- Computing R&D projects started in the mid-90s – (others much earlier)
- No explicit focus on preservation at this stage, rather on Mass Storage Systems, Data Formats and I/O systems
- This included IBM's HPSS, Objectivity/DB and a major move from Fortran to C++
- C++ has remained, but both HPSS and Objectivity/DB were replaced:

➢ **Major data migrations requiring new s/w**

2. Enter the Grid (A Y2K Bug?)

- (W)LCG based on a hierarchical model of O(1) Tier0s, O(10) Tier1s and O(100) Tier2s
  - Sum of resources at each level ~constant
  - Service levels decrease with increasing Tier #

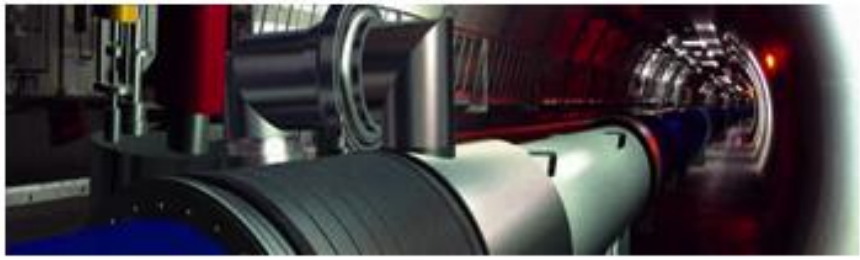
- **"Data curation" – a responsibility of the Tier0 and Tier1 centres (LCG TDR 2005)**
  - *"A Tier-1 is responsible for the storing and general curation of archived data through the life of the experiments it supports.*
  - *This implies the need to retain capabilities in the technology in which the data is originally recorded or to migrate the data to new storage technologies as they are adopted at that site in the future."*

➢ **This was the first time (AFAIK) data preservation was considered prior to data taking**
  - Although clearly not thought out in much detail, nor costed

LCG Service Hierarchy

Tier-0 – the accelerator centre

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data → Tier-1 centres



Tier-1 – "online" to the data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- Data-heavy analysis
- National, regional support

Tier-2 – ~100 centres in ~40 countries

- Simulation
- End-user analysis – batch and interactive

2. The DPHEP Blueprint

- Towards the end of “the noughties” (2000 decade), a number of Collider experiments were coming to an end worldwide
- Together, formed the DPHEP Study Group that produced a Blueprint document outlining the motivation for LTDP in HEP as well as major issues (funding, support)

➢ **Fed into the 2012 ESPP update and resulted in LTDP being a part of the current strategy**

The DPHEP Blueprint

- **Comprehensive review of the HEP situation**
- **Action plan proposed for the next 3 years**

- **A cry for help:**

  - ***Urgent action is needed for LTDP in HEP***
  - ***The preservation of the full capacity to do analysis is recommended such that new scientific output is made possible using the archived data***

2013 ESPP on LTDP

- *…**data preservation** and distributed data-intensive computing **should be maintained and further developed**.*

- **Well – at least it got a mention!**

- **Hope to have something a bit more concrete next time! (2030 vision?)**

2. From Study Group to Collaboration

- **Primary goal** was to **deliver solutions** to the issues outlined in the Blueprint

1. Preparing a 2020 Vision (shown many times)
2. Understanding the Full Costs of Curation
3. Agreeing HEP-wide on the core services on which LTDP depends

   a. "Bit **preservation**" (of the data) **(?? TDRs ??)**
   b. Documentation **preservation** (more complex than just a few PDFs)
   c. **Preservation** of software + **environment in which it runs**

   VERY DIFFERENT to porting s/w forward…

   - Such services in production since 2015 at CERN, outside and in EOSC*
   - **Although I don't believe in generic TDRs… See later…**

➢ **See https://indico.cern.ch/category/4458/**

LTDP in HEP

iPRES 2016
13th International Conference on Digital Preservation //
Bern // October 3–6, 2016

- Built on 3 preservation "pillars":
  1. **The data itself** ("**bits**" – state of the art bit preservation);
  2. **Documentation** (services like Zenodo, B2SHARE);

together with the necessary

3. **Software + environment (CernVM / CVMFS)**

- Services for all 3 areas **exist** and are **mature** but **change** on fully independent timescales
- We need flexible (not static) bridges between them

3. Future Directions

1. “Analysis preservation” and reproducibility of results are top priorities for Data Producers
2. Solving the Open Data funding / resource problem
3. Data Preservation for non-CERN experiments
4. Completing ISO 16363 certification seen as important for CERN, possibly also ESFRIs + ?
5. Tape alternatives? Concerns over size of Enterprise Tape market (we are a small player)
6. Use of commercial or other Cloud Backends?
7. PV2020 and “EIROforum” meetings

3. Future Directions (I)

1. "Analysis preservation" and reproducibility of results are top priorities for Data Producers
   - This could be highly complex but experiments "agree" on what information must be captured; prototyping how to do it
   - Possible RIA project (January 2019) in this area
2. Solving the Open Data funding / resource problem
   - My proposed solution: define what resources possible and cache only recent / "featured" / requested data (NOT ALL!)
3. Data Preservation for non-CERN experiments
   - By example: WLCG / HSF workshop at JLAB March 2019
   - Ingest & preservation of 2PB BaBar data (SLAC)

Topic: INFRAEOSC-02-2019

Title: Prototyping new innovative services

Challenge:

Develop an agile, fit-for-purpose and sustainable service offering accessible through the EOSC hub that can satisfy the evolving needs of the scientific community by stimulating the design and prototyping of novel innovative digital services. Innovative models of collaboration that genuinely include incentive mechanisms for a user oriented open science approach should be considered.

Scope:

- Actions that **target gaps in the service offering of the EOSC hub** and develop innovative services that address relevant aspects of the research data cycle (**from inception to publication, curation, preservation and reuse**), e.g. allowing implementation of new scientific data-related developments and intelligent linking and discovering of all research artefacts.
- By the end of the project the services should be leveraged to **foster interdisciplinary research**, serving a wider remit of research needs, as well as new users like industry and the public sector.
- The services should be based on systems and technologies that have reached **TRL 6 b**efore the start of the project and will be brought to **at least TRL 8 by the end** of the project.
- Proposals should demonstrate how the resulting services **complement, enrich and could potentially be integrated into the EOSC hub.**
- Proposals retained for funding under this topic should take due consideration of any accessibility requirements set under the projects funded under EINFRA-12-2017 topic that may be available at the time the call will be open.
- Consortia are encouraged to include **SMEs**

Type of Action: **RIA**

Budget: **€28.5M**

Submission opening: **16/10/2018**

Submission deadline: **29/01/2019**

Suggested contribution requested per proposal: **€5-6M**

3. Future Directions (II)

4. Completing ISO 16363 certification seen as important for CERN, possibly also ESFRIs + ?
   - Self-certification drafted; Stage 1 “off-site” audit performed; planning formal on-site audit
   - N.B. covers all LTDP activities at CERN – not just scientific data
   - Certification of some ESFRIs as part of ESCAPE project

5. Tape alternatives? Concerns over size of Enterprise Tape market (we are a small player)
   - Possible topic for EIROForum+ TWG meetings
   - CERN is following this internally, with Tier1s and big players such as Internet giants

6. Use of commercial or other Cloud Backends?
   - H2020 PCP ARCHIVER project (January 2019) will prototype this at 100TB – 1PB level
   - Can we agree on common Use Cases and interface(s)?

3. Future Directions (III)

7. PV2020 and “EIROforum” meetings

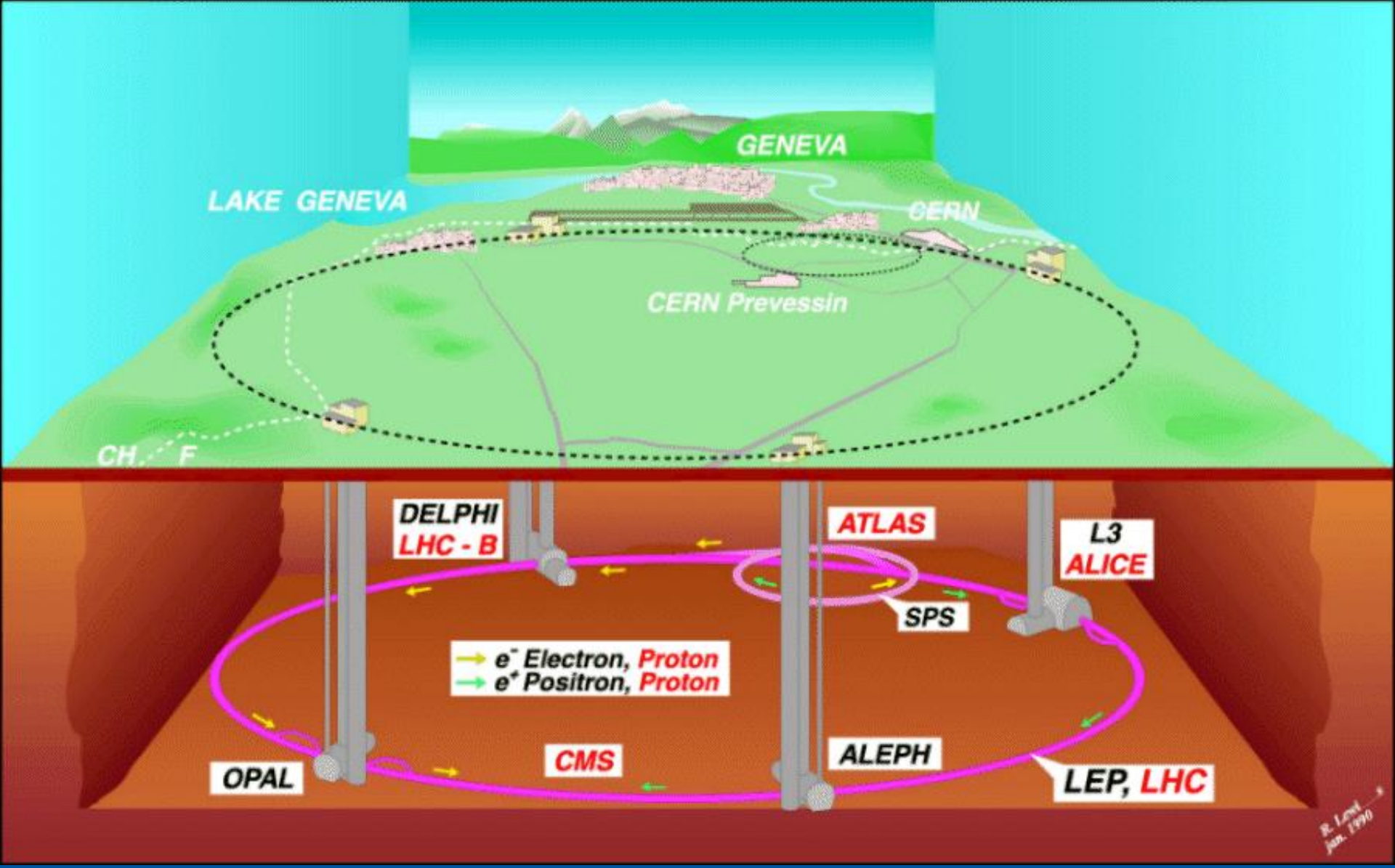
- Have discussed holding short, focused technical meetings with EIROforum (and other) members
- Tapes and tape alternatives is a possible 1st topic
- Interest?

- PV2020 tentative dates and timeline established
- Still much to do – including discussing with Esther!

Summary

- Sustainable data preservation is a reality at CERN: **300PB** with an outlook to **10EB**
- Now addressing “next level” concerns, including reproducibility of results, **Open Data at multi-PB** level, potential use of (commercial) cloud storage etc.
- Hopefully, we can strengthen our links, arrange regular technical meetings …

➢ **And make PV2020 “Simply the Best!”**



LEP Experiments: ALEPH, DELPHI, L3, OPAL (e^+e^-)

LHC Experiments: ALICE, ATLAS, CMS, LHCb (pp + HI)

Overview

- 2012 DPHEP “Blueprint” published & input to European Strategy of Particle Physics update
  - Strategy adopted by CERN Council May 2013
- 2018: a new strategy update has been launched with input due December 2018
  - To be adopted in (May?) 2020

1. **What was the situation in 2012/13?**
2. **What have we achieved since then?**
3. **Outstanding issues & future directions**

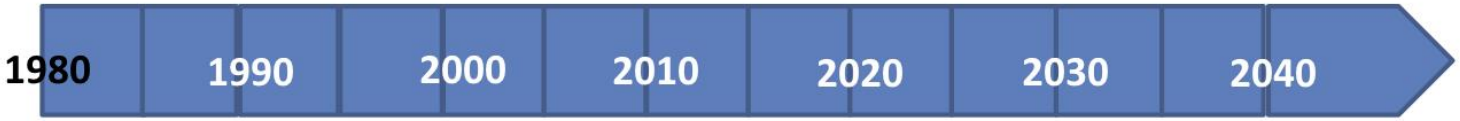
During this period the world about us has changed considerably (4C, FAIR DMPs, EOSC*, etc.)

Definitions

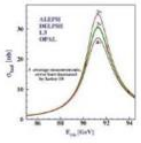
- "Long-term" in HEP means "a few decades"
  - **LEP started in 1989 – data alive until 2030?**
    - **~100TB** per experiment (ALEPH, DELPHI, L3, OPAL)
  - **LHC started in 2009 – upgrades until 2035+**
    - Also 4 main experiments (ALICE, ATLAS, CMS, LHCb)
    - Already in excess of **200PB**

- "Collaborative" means
  1. Across all HEP institutes and experiments
  2. Together with other disciplines / institutes (and in particular ESFRIs and EIROForum)
  3. With Funding Agencies, Policy Makers, EOSC & other projects and also Industry

LEP / (HL-)LHC Timeline

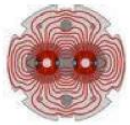


LEP Constr. | Physics | Upgr.



Database / data management support,
CERN Program Library, Distributed Computing

LHC Design, R&D | Proto. | Constr. | Physics



DM R&D, DBs, WLCG, EGI
Major Data Migrations(!)

HL-LHC Design, R&D | Constr. | Physics

High Luminosity LHC

ESFRI roadmap as "landmark project"

- Robust, stable services over **several decades**
- Data preservation and re-use over **similar periods**
- "Transparent" and supported **migrations**

What does DPHEP do?

- DPHEP is **a Collaboration** with signatures from the main HEP laboratories and some funding agencies **worldwide**.

- It has established a "**2020 vision**", whereby:
  - All archived data – e.g. that described in DPHEP Blueprint, including LHC data – should be easily **findable** and fully usable by the **designated communities** with clear (Open) access policies and possibilities to annotate further;
  - Best practices, tools and services should be well run-in, **fully documented** and **sustainable**; built in common with other disciplines, based on standards;
  - There should be a DPHEP **portal**, through which data / tools accessed;
  - Clear **targets & metrics** to measure the above should be agreed between Funding Agencies, Service Providers and the Experiments.

The DPHEP Blueprint

- **Comprehensive review of the HEP situation**
- **Action plan proposed for the next 3 years**

- **A cry for help:**

  - ***Urgent action is needed for LTDP in HEP***
  - ***The preservation of the full capacity to do analysis is recommended such that new scientific output is made possible using the archived data***

2013 ESPP on LTDP

- *…**data preservation** and distributed data-intensive computing **should be maintained and further developed**.*

- **Well – at least it got a mention!**

- **Hope to have something a bit more concrete next time! (2030 vision?)**

LTDP in HEP

iPRES 2016
13th International Conference on Digital Preservation //
Bern // October 3–6, 2016

- Built on 3 "pillars":
  1. **The data itself** ("**bits**" – state of the art bit preservation);
  2. **Documentation** (services like Zenodo, B2SHARE);

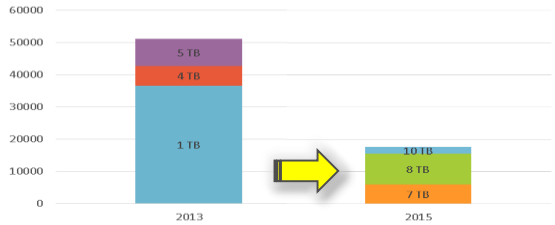
  together with the necessary

  3. **Software + environment (CernVM / CVMFS)**

- Services for all 3 areas **exist** and are **mature** but **change** on fully independent timescales
- We need flexible (not static) bridges between them

Bit Preservation: Steps Include

- Regular media **verification**
  - When tape written, filled, every 2 years…
- Controlled media **lifecycle**
  - Media kept for 2 max. 2 drive generations
- **Reducing** tape mounts
  - Reduces media wear-out & increases efficiency
- Data **Redundancy**
  - For “smaller” communities, a 2nd copy can be created: separate library in a different building (e.g. LEP – **3 copies at CERN**!)
- **Protecting** the physical link
  - Between disk caches and tape servers
- Protecting the **environment**
  - Dust sensors! (Don’t let users touch tapes)



Constant improvement: reduction in bit-loss rate: 5 x 10^{-16}

Documentation - Some Issues

- **Some preservation systems claim to do automatic format conversion (of documents)**
- **IMHO this is not realistic for e.g. scientific documentation**
  - Maybe for basic "ASCII text"
    - but see e.g. Kindle format books
  - 1018 and 10^{18} are not the same – let alone formulae for Bessel functions of the 3rd kind!
- **2016 re-formatting of CERNLIB docs required detailed <u>mathematical</u> knowledge in addition to reprocessing LaTeX into PDF!**

What is HEP data?

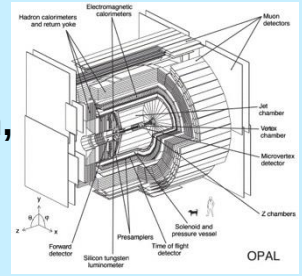
Digital information

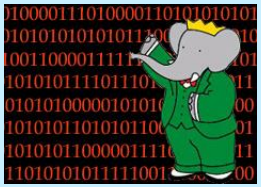
The data themselves, volume estimates for preservation data of the order of a few to 10 PB (to 10 EB for LHC)

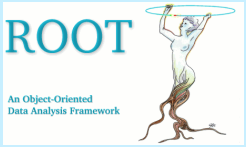
Other digital sources such as databases to also be considered

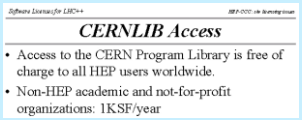
Software

Simulation, reconstruction, analysis, user, in addition to any external dependencies



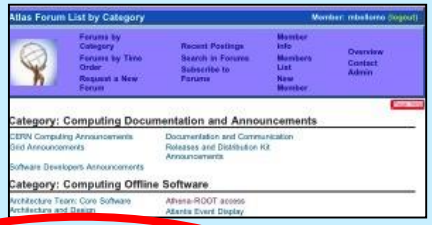






Meta information

Hyper-news, messages, wikis, user forums..



Publications

arXiv.org

INSPIRE

Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

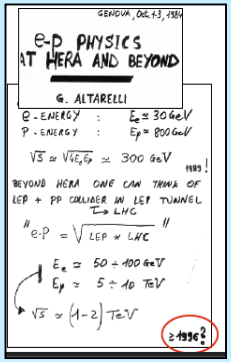
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,) and is updated at regular intervals.

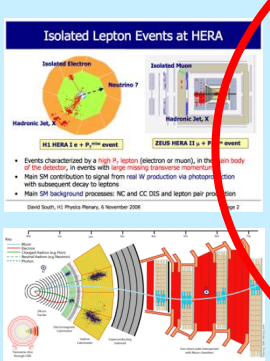
Journal of High Energy Physics

A refereed journal, written, run and distributed by electronic means

Documentation

Internal publications, notes, manuals, slides





Expertise and people

DESY

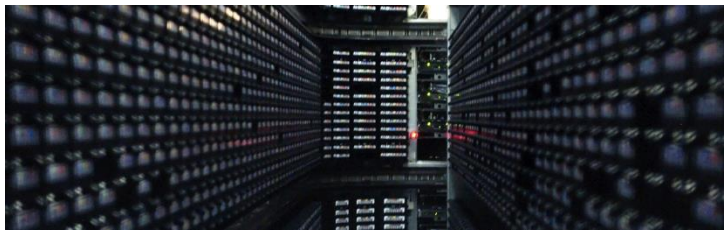
Software Preservation

- Today HEP s/w is O(10^7) lines of code, 10s to 100s of modules and many languages! (**"Bespoke" per experiment AND analysis**)
- ➢ **Versioning filesystems and virtualisation** widely adopted in HEP: also in EGI & EOSC!
- Capture and preserve software + environment for **each set** of preserved data
- ➢ Believe we can analyse LEP data **~30 years** after data taking ended! (i.e. until ~2030)

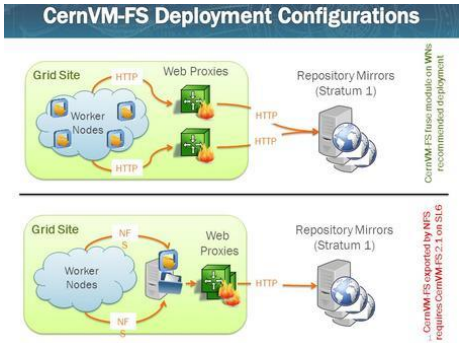
F.A.I.R. Data Management

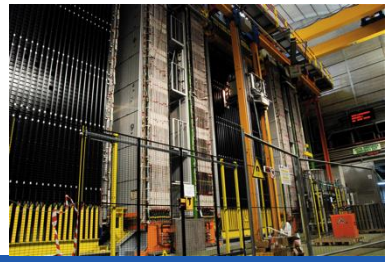
- Increasing emphasis on FAIR DMPs, including preservation, sharing, reproducibility etc.
  - **FAIR now includes also s/w but not yet build systems, verification procedures & environment**
- IMHO not yet fully understood (some claim otherwise) - we see (ir)regular changes on how we **find** data and what protocol(s) we use to **access** it
- **This can be a problem over periods < 1 decade**
- **Only solution we know of: find the effort to migrate (problem for legacy projects / data)**

Collaboration



1. **Through technology:**
   - Large Tape Users' Group
   - Invenio → Zenodo → B2SHARE (INSPIREHEP)
   - CVMFS / CernVM
2. **Through projects:**
   - e*, E* and H*
3. **Through services:**
   - Obi-wan Zenodo
   - CVMFS repository for "lost" experiments
   - Possible hosting of 2PB of BaBar@SLAC data
   - 70 TB of OPERA data (CERN "recognised" expt)
4. **Through workshops & conferences:**
   - e.g. EIROForum technical WG on LTDP
   - More on Thursday… (?)





BABAR needs Help!

- *BABAR* data actively being analyzed and high impact papers published (see slide 2). Expect this to continue to at least through 2021.
- SLAC management plans to **stop** hosting *BABAR* computing in February 2020 at which time the tapes with **data will be ejected**.
- DOE support ended in 2017, now running on international common funds (OCF).
- Looking for possibility of support and long term data preservation at
  - CERN,
  - GridKa (*BABAR* site for analysis and XRootD federated dataset main redirector),
  - University of Victoria (*BABAR* site for analysis, documentation, and tools support).
- *BABAR* lightweight VMs come with the latest software release and xrootd client included, running under the most common virtual machine players. Just add the data via the GridKa main XRootD redirector.

BABAR in Numbers

- **2PB of data on T10k-D tapes**
  - raw, processed, Monte Carlo
  - **Unique dataset at the Y(3S) resonance** (no plan (yet?) to run at the Y(3S) @ Belle II)
- Full environment enclosed in VMs (SL5,SL6)
- ~1TB of documentation, repositories, and dataset information (DBs, cvs, wiki, html)
  - Internal documents archived on INSPIRE

- 574 papers, ~10 papers/year past 3 years
- 231 members (semi-frozen author list)
  - Including PhD students in Canada, Germany, Israel, Italy, Russia, US
  - Associated theorists mine data to test new ideas
- ~20 analyses on track, ~10 more in the pipeline
  - Continue to have new analyses every year including joint *BABAR* -Belle analyses
- **Students analyze *BABAR* data while working on Belle II and other experiments in construction/commissioning phase**

ISO 16363 certification of CERN

- ISO 16363 follows OAIS breakdown:
  3. **Organisational Infrastructure;**
  4. **Digital Object Management;**
  5. **Infrastructure and Security Risk Management.**

- Many of the elements in 3) and 5) covered by existing (and documented) CERN practices
  - **Some "weak" areas – being addressed – include disaster preparedness / recovery (together with EIROForum)**

➢ **On-going "stage 1" external audit to high-light those areas requiring attention**
  - **May just be a question of documentation, e.g. CERN is not going to change its financial practices (MTP etc) as a result of ISO 16363!**

Who does it benefit?

- **Funding agencies**, who can better judge if the money they are providing will be used according to their requirements
  - e.g**. FAIR DMPs which call for preservation & re-use**
- **Data users** to be able to determine the "trustworthiness" of the data (**user surveys**)
- **Producers** (e.g. LHC experiments) to understand how and what a repository does to preserve their data

➢ **The data of most CERN experiments already lost!**

- **By number of experiments, not by volume**
  - CERN Greybook: 776 completed experiments, ~20 active
  - "Preserved": LEP(4), LHC(4)
  - ➢ **O(10) vs O(1000)**

HEP Community White Paper

- Focuses on the challenges of the next decade or so (LHC Run3, HL-LHC Run4)
  - **Massive increase in data rates and computational needs – way beyond technology predictions**
- ***"bit preservation with an acceptably low error rate can now be considered a solved problem"***
- Main areas of work now:
  - Analysis capture (incl. workflows) and reproducibility
  - "Open Data" at **multi-PB** scale and beyond
  - Trying to do this in collaboration with others (e.g. RDA)
- **Does "Open Data" mean zero or low latency?**
  - **People assume so – enormous implications!**

Services are (just) services

- No matter how fantastic our { TDRs, PID services, Digital Library, Software repository } etc is, they are there to support **the users**

➢ **Who have to do the <u>really</u> hard work!**

- **E.g. write the software, documentation, acquire and analyse the data, write the scientific papers**

- However, getting the degree of public recognition as at the Higgs discovery day was a **target e-KPI**!

What is the future?

- Some hope that it may be possible to separate long-term preservation of data at the **bit level** from **domain-specific aspects**
- **The former could benefit from economies of scale and specialised knowledge in running multi-PB / EB archives**
- ➢ **The latter will continue to need expert knowledge to revalidate on a regular basis**
- Drive to reduce overhead through "domain protocols" for DMPs

Input to next ESPP

- Certification as a “Very Trustworthy Digital Repository” – exabytes & decades & changes
- Open Data – clarification(s); resources
- Reproducibility & Re-use
- **Resilience to and handling of change(s)**

29 years of LEP – what does it tell us?

- Major migrations are **unavoidable** but hard to **foresee**!
- **Data** is not just "**bits**", but also **documentation, software + environment + "knowledge"**
  - "**Collective knowledge**" particularly hard to capture
    - Documentation "refreshed" after 20 years (1995) – now in Digital Library in PDF & PDF/A formats (was Postscript)
- Today's "**Big Data**" may become tomorrow's "**peanuts**"
  - 100TB per LEP experiment: **immensely challenging** at the time; now "trivial" for both CPU and storage
  - With time, **hardware costs** tend to zero
    - O(CHF 1000) per experiment per year for archive storage
  - **Personnel costs** tend to O(1FTE) **>> CHF 1000!**
    - Perhaps as little now as 0.1 – 0.2 FTE per LEP experiment to keep data + s/w alive – (new analyses "cost extra")

Conclusions

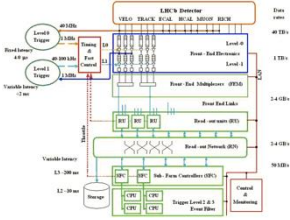
- We are well on the way to implementing our 2020 vision using “standard” services
  - VTDR & PIDs, Digital Libraries & DOIs, s/w preservation
  - Services that are – or should be – offered in the EOSC*

- **But they are not “holistic” – “mind the gap(s)”**
- **And they will change over time – whatever people (especially in IT) pretend!**
- **Beware of "grey-backed gorillas"**

==> Constant effort is needed – like with a bike


CERN

What Makes HEP Different?

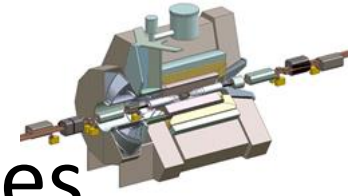


- We **throw away** most of our data before it is even recorded – "triggers"
- Our detectors are **relatively stable** over long periods of time (years) – not "doubling every 6 or 18 months"
- We make "**measurements**" – not "**observations**"
- Our projects typically last for **decades** – we **need** to keep data usable during at least this length of time **(but not necessarily "forever")**
- We have **shared** "data behind publications" for more than 30 years… (HEPData)

ODBMS migration – overview (300TB)

- **A triple migration!**
  - Data format and software conversion from Objectivity/DB to Oracle
  - Physical media migration from StorageTek 9940A to 9940B tapes
- **Took ~1 year to prepare; ~1 year to execute**
- **Could never have been achieved without extensive system, database and application support!**
- Two experiments – many software packages and data sets
  - **COMPASS** raw event data (300 TB)
    - Data taking continued after the migration, using the new Oracle software
  - **HARP** raw event data (30 TB), event collections and conditions data
    - Data taking stopped in 2002, no need to port event writing infrastructure
  - In both cases, the migration was during the "lifetime" of the experiment
  - System integration tests validating read-back from the new storage

BABAR Highlights and Press Releases



BERKELEY LAB
Bringing Science Solutions to the World

About the Lab | Leadership/Organization | Calendar | News Cent

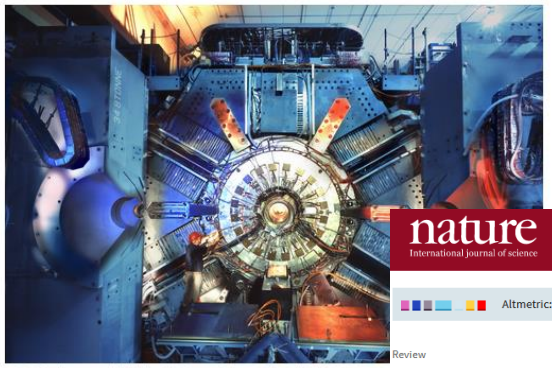
NEWS CENTER

New Study: Scientists Narrow Down the Search for Dark Photons Using Decade-Old Particle Collider Data

Analysis of data from the BaBar experiment rules out theorized particle's explanation for muon mystery

News Release Glenn Roberts Jr. (510) 5

November 2017



The BaBar detector at SLAC National Accelerator Laboratory. (Credit: SLAC)

Dataset:
Y(4S): 433/fb
Y(3S): 30/fb
Y(2S): 14/fb
Off resonance: 10%
Y(1S) accessed via
$\Upsilon(2S,3S) \rightarrow \Upsilon(1S)\,\pi^+\pi^-$

INTERACTIONS.ORG
PARTICLE PHYSICS NEWS AND RESOURCES

Home | About | Particle People

A communication resource from the **world's particle physics laboratorie**

New study: Scientists narrow down the sea photons using decade-old particle collider d

8 November 2017 - Lawrence Berkeley National Laboratory

Analysis of data from the BaBar experiment rules out theorized particle's explanation for muon mystery

In its final years of operation, a particle collider in Northern California was refocused to search for signs of new particles that might help fill in some big blanks in our understanding of the universe.

A fresh analysis of this data, co-led by physicists at the Department of Energy's Lawrence Berkeley National Laboratory (Berkeley Lab), limits some of the hiding places for one type of theorized particle – the dark photon, also known as the heavy photon – that was proposed to help explain the mystery of dark matter.

The latest result, published in the journal *Physical Review Letters* by the roughly 240-member BaBar Collaboration, adds to results from a collection of previous experiments seeking, but not yet finding, the theorized dark photons.

"Although it does not rule out the existence of dark phot hide, and definitively rule out their explanation for anothe property of the subatomic particle known as the muon," University of Victoria professor.

Dark matter, which accounts for an estimated 85 percent been observed by its gravitational interactions with norm galaxies is much faster than expected based on their visi

DATE ISS
November 8
SOURCE
Lawrence B
CONTEN
Press Release
CONTACT:
y National Labo

nature
International journal of science

Altmetric: 512 | Citations: 2 | More deta

Review

A challenge to lepton universality in B-meson decays

Gregory Ciezarek[1], Manuel Franco Sevilla[2], Brian Hamilton[3], Robert Kowalewski[4], Thomas Kuhr, Vera Lüth & Yutaro Sato[7]

Nature **546**, 227–233 (08 June 2017) | Received: 15 December 2016
doi:10.1038/nature2 | arch 2017
Download Citation | e: 07 June 2017

June 2017

Experimental particle physics | Phenomenology | Theoretical particle physics

Abstract

Physics ABOUT BROWSE PRESS COLLECTIONS

Viewpoint: New Light Shed on Dark Photons

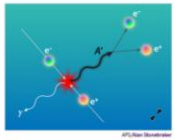
Douglas Bryman, University of British Columbia, Vancouver, British Columbia V6T2A3, Canada
November 10, 2014 • Physics 7, 115

A search for a photonlike particle that could be related to dark matter has come up empty, putting new constraints on models that imagine a dark form of electromagnetism.

PDF Version | Print

Search for a Dark Photon in e^+e^- Collisions at *BaBar*
J. P. Lees *et al.* (BaBar Collaboration)
Phys. Rev. Lett. 113, 201801 (2014)
Published November 10, 2014

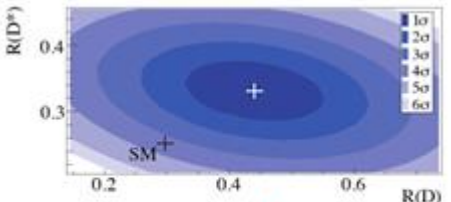
Read PDF

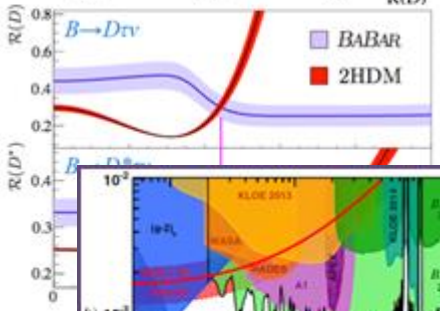


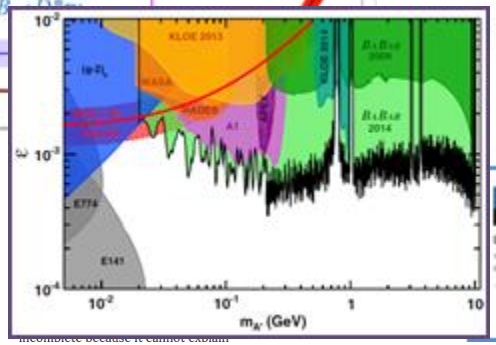
Features

Q&A: A Condensed Matter Theorist Embraces AI
Juan Carrasquilla gave himself a crash course on machine learning and found a new way of approaching condensed-matter theory.

Meetings: Interplanetary GPS





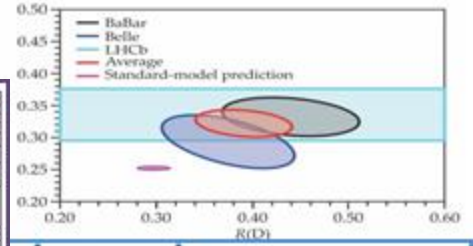


The Economist

The arrow of time
Backward ran sentences...
To the relief of physicists, time really does have a preferred direction

TIME seems to flow inexorably in one direction. Superficially, that is because things deteriorate with age—and this, in turn, is because there are innumerably fewer ways to arrange particles in an orderly fashion than in a jumbled mess. Any change in an existing arrangement is therefore likely to increase its disorder.

Dig a little deeper, though, and time's arrow becomes mysterious. A particle cannot, by itself, become disordered, so when you examine its behaviour in isolation the past and the future are hard to distinguish. If you film its movement and then give the film to someone else, he will not be able to work out just from the particle's behaviour which way to run the film through the projector.



physicstoday

Democracy suffers a blow—in particle physics