IDKM in High Energy Physics

Jamie.Shiers@cern.ch

https://indico.cern.ch/event/791540/

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

Introduction

- Normally when I present LTDP activities at CERN and in HEP in general, I do so in terms of existing, production services.
- ➢ **But such an approach hides the fact that the situation is fragile and – at least looking back to LEP, we arrived here almost by accident!**
- **Furthermore, there have been several developments in the past year that are worth emphasizing**
  - H2020 projects (ARCHIVER, ESCAPE, …)
  - Collaboration (DPC and others, …)
  - Conferences (iPRES 2019, PV 2020@CERN, …)
  - ➢ **Tentative Schedule for FCC published!**

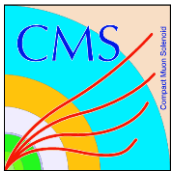
Technical aspects of IDKM

- Short background of the (worldwide) HEP community
- Why IDKM is important for us
- Global vision and plans for an efficient IDKM
- Examples of past and on-going projects in IDKM
- What are the main lessons learnt about IDKM
- Future plans
- Suggestions for the RWM scientific community

Once upon a time (end 2008)

- Lots of HEP data collected, and lots of data to come
  - PEP2,HERA,Tevatron shutdown
  - The LHC start-up
- Some experiments wondered what was going to happen to their data
  - Is it going to be relevant?
  - Was it prepared?
  - Costs? Personpower?
  - Coordination?
- Study group convened: contacts to many labs and experiments, decided to focus on large colliders
  - Issue at ICFA meeting in SLAC in Nov 2008

DPHEP: An international study group on data preservation


Institute of High Energy Physics
Chinese Academy of Sciences



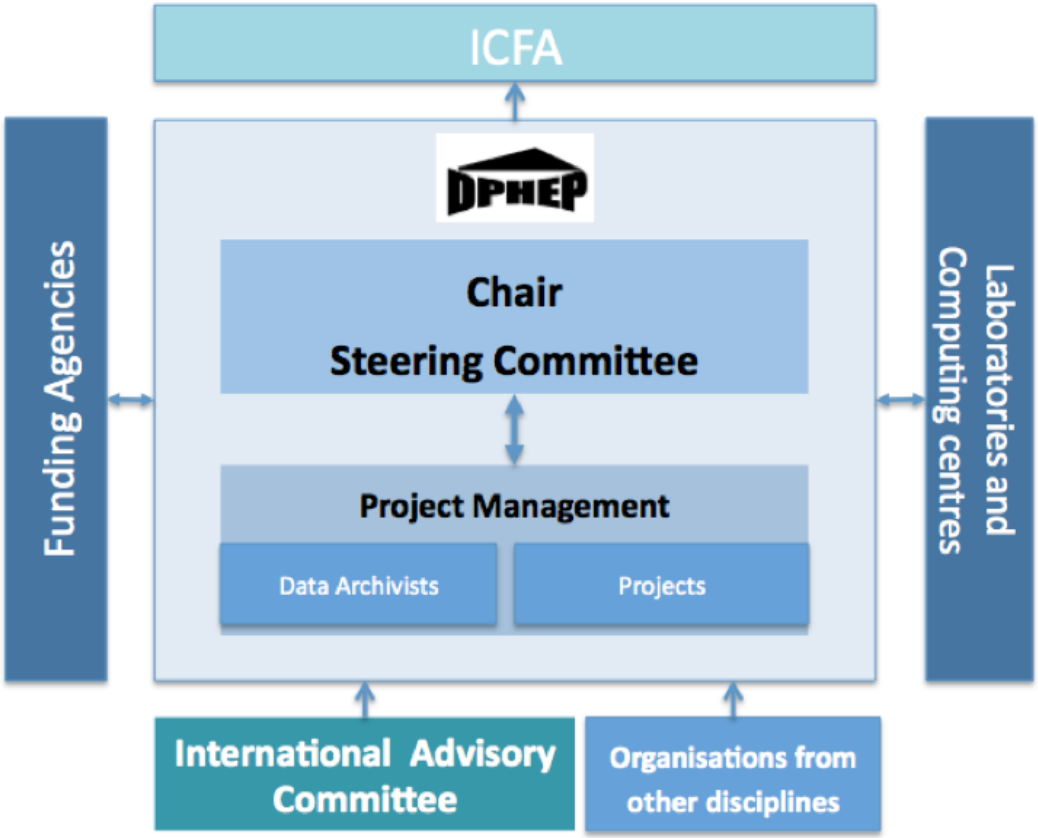
Science & Technology
Facilities Council


DPHEP Blueprint Messages

- DPHEP Blueprint published in **May 2012** – to some extent 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***
- Current ESPP (**May 2013**):
  - *…**data preservation** and distributed data-intensive computing **should be maintained and further developed**.*

The DPHEP Organisation (Long-Term Sustainability)

- Retain the basic structure of the Study Group, with links to the host experiments, labs, funding agencies, ICFA
- **Installation of a full time DPHEP Project Manager, who acts as the main operational coordinator**
- The DPHEP Chair (appointed by ICFA) coordinates the steering committee and represents DPHEP in relations with other bodies



Not as successful in HEP as e.g. "Space"

What does DPHEP do?

- DPHEP has become **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.

CERN Services for LTDP

1. State-of-the art "**bit preservation**", implementing practices that conform to the ISO 16363 standard
2. "**Software preservation**" - a key challenge in HEP where the software stacks are both large and complex (and dynamic)
3. Analysis **capture and preservation**, corresponding to a set of agreed Use Cases
4. Access to **data behind physics publications** - the HEPData portal
5. An **Open Data portal** for released subsets of the (currently) LHC data
6. A **DPHEP portal** that links also to data preservation efforts at other HEP institutes worldwide.

➢ **Each of these is a talk topic in its own right!**

Certification of the CERN Site

- We believe **(re-)**certification will allow us to ensure that best practices are implemented and followed up on in the long-term: "**written into fabric of organisation**"

- Scope: **Scientific Data** and CERN's **Digital Memory**

- Timescale: complete prior to 2019/2020 ESPP update
- Will also "ensure" adequate resources, staffing, training, succession plans etc.

- CERN can expect to exist until HL/HE LHC (2040/50)
- And beyond? FCC? Strategy decisions 2020 / 2030

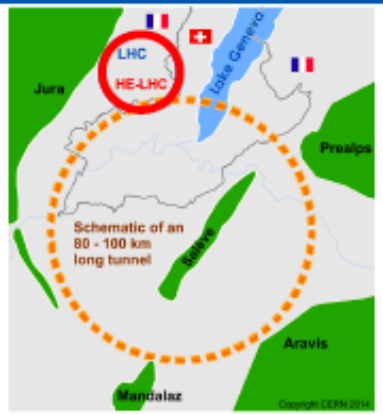
Overview of CERN

- CERN – the European Organisation for Nuclear Research – is situated just outside Geneva, extending into France
- Founded in **1954**, it now has **22** member states
- **It operates a wide range of accelerators of which the LHC is probably best known**
- *A* **L**arge **H**adron **C**ollider was first proposed in the late 1970s, when discussions on a Lepton Collider (**LEP**) were being held
- A High Luminosity upgrade (**HL-LHC**) was approved in June, extending the LHC's life until around 2040
- High Energy (**HE-LHC**) & even **FCC** may follow…

Future Circular Collider (FCC)

Conceptual Design Report released today!

	√s	L /IP ($cm^{-2} s^{-1}$)	Int. L /IP(ab^{-1})	Comments
e^+e^- FCC-ee	~90 GeV Z 160 WW 240 H ~365 top	230 x10^{34} 28 8.5 1.5	75 ab^{-1} 5 2.5 0.8	2 experiments Total ~ 15 years of operation
pp FCC-hh	100 TeV	5 x 10^{34} 30	2.5 ab^{-1} 15	2+2 experiments Total ~ 25 years of operation
PbPb FCC-hh	$\sqrt{s_{NN}}$ = 39TeV	3 x 10^{29}	100 nb^{-1}/run	1 run = 1 month operation
ep Fcc-eh	3.5 TeV	1.5 10^{34}	2 ab^{-1}	60 GeV e- from ERL Concurrent operation with pp for ~ 20 years
e-Pb Fcc-eh	$\sqrt{s_{eN}}$ = 2.2 TeV	0.5 10^{34}	1 fb^{-1}	60 GeV e- from ERL Concurrent operation with PbPb



Also studied: HE-LHC: √s=27 TeV using FCC-hh 16 T magnets in LHC tunnel; L~1.6x10^{35} → 15 ab^{-1} for 20 years operation

Sequential implementation, FCC-ee followed by FCC-hh, would enable:

- ❑ variety of collisions (ee, pp, PbPb, eh) → impressive breadth of programme, 6++ experiments
- ❑ exploiting synergies by combining complementary physics reach and information of different colliders → maximise indirect and direct discovery potential for new physics
- ❑ starting with technologically ready machine (FCC-ee); developing in parallel best technology (e.g. HTS magnets) for highest pp energy (100++ TeV!)
- ❑ building stepwise at each stage on existing accelerator complex and technical infrastructure

Purely technical schedule, assuming green light to preparation work in 2020.

A 70 years programme

8 years preparation	10 years tunnel and FCC-ee construction	15 years FCC-ee operation	11 years FCC-hh preparation and installation	25 years FCC-hh operation pp/PbPb/eh
2020-2028		2038-2053		2064-2090

LTDP & FCC Options

- LEP (1989 – 2000) vs FCC ee (2038 – 2053)
- HERA (1992 – 2007) vs FCC ep (2064 – 2090)
- LHC (2008 – 2035) vs FCC hh (2064 – 2090)

- Can we keep LEP data alive until ~2040?
  - **Best guess currently 2030**

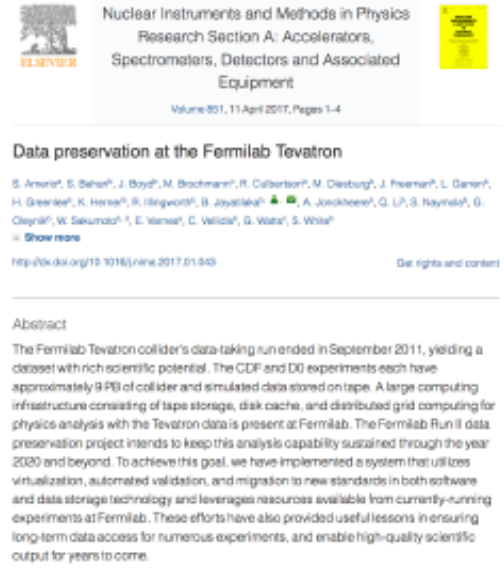
- Can we keep HERA data alive until ~2065?
  - **Best guess currently 2020**

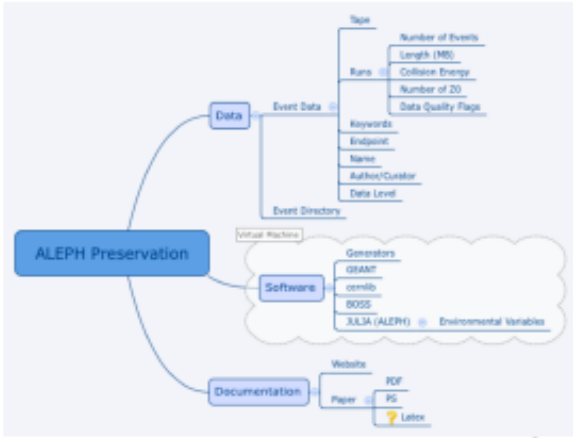
- Can we keep LHC data alive until ~2065?

Data Preservation projects in pre-LHC experiments

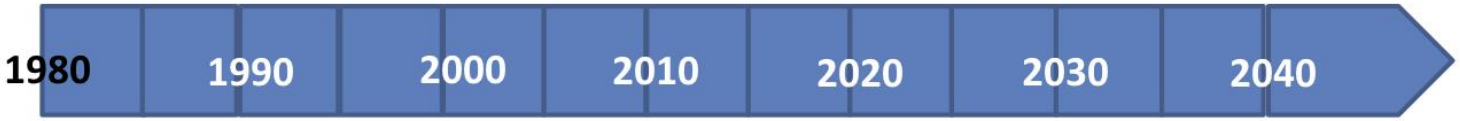
- **@Babar:** LTDAP system central to the scientific output since 2012 (getting old), objective to keep alive by 2020.
- **@DESY:** H1 (migration) and ZEUS (encapsulation) successful transition to the DP systems , publication plan includes ~20 more papers, alive by 2020
- **@FNAL**: recent publication, successful transition to a DP system for both CDF (CDFDP) and D0 (R2DP)
- **@KEK:** BELLE I data readable in Belle II framework (2020)
- **@IHEP/BES3:** The experiment is expected to stop data taking at 2022, and Lifespan of preserved data is expected to be about 15 years after then.
- **@CERN**: ALEPH, DELPHI, OPAL data/sw refreshed, OPERA; central work on technology, DPHEP portal
- **@MPI**: multi-experiment framework explored (JADE, HERA, OPAL)

- Transition from experiment-driven to lab-supported systems successfully accomplished in general
  - Physics output continues on new "DP"-like systems
- Clear need to keep the issue highly visible on the community's agenda
  - ensure an adequate level of endorsement from FA/Labs/Experiments

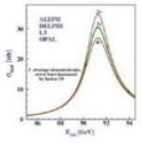




LEP / (HL-)LHC Timeline

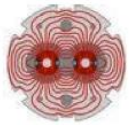


LEP Constr. | Physics | Upgr.



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

ESFRI roadmap as "landmark project"

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

Data Preservation at CERN

- Concerns not just scientific data but also papers and publications (stored in a Digital Library with DOIs) as well as CERN's "Digital Memory"
  - Photos, videos, minutes etc. (many already lost...)
- Same basic technologies and services but VERY different impact if data lost / degraded

- **LEP machine: dismantled post 2000 and location of each component stored in a database**

- **How long is "forever?"**

- ***Until I get a new job / retire / lose interest?***

RETIREMENT WILL END SHIERS' LONG RUN AT CERN

An OF physicist of the highest calibre is about to retire. Here he tells us about CERN and his future. Joel Garner writes

One of Felsted's most-respected physicists Jamie Shiers (fd66-74) will retire in three years' time ending a lifetime's work at CERN, the European Organisation for Nuclear Research.

His retirement in 2021 will end his long association with the Swiss-based research centre in Geneva, which provides particle accelerators and infrastructure needed for high-energy physics research.

"I'd say I've given what I can to CERN," he said, before adding it was time for a younger generation to take the organisation forward. Retirement will also mean that he can spend more time with his British wife Samantha and young children, Lancelot, who is seven years old and daughter Kalinka who is five. This may also mean being able to enjoy his hobbies, sailing and skiing and complete some unfinished business.

"I have a project to finish the 10 Chopin waltzes where only the first line still exists – the complete scores were lost in a fire," he added.

By coincidence his retirement starts at the end of the second long shut down of the Large Hadron Collider (LHC) and as it prepares for its ambitious performance upgrade that will mean the scientists will be able to observe more rare phenomena and study other known ones in more detail.

Shiers, who is data preservation manager at CERN, created and maintained the current LHC computing grid, the world's largest, to collect, store and analyse the data from high-energy experiment and share it with the world.

Despite working in the IT department, he is a physicist at heart and has maintained his love for the subject which he has always found fascinating. He added that his knowledge of physics has been vital to his role. "If I didn't have a good understanding of what the scientists are trying to do it would be impossible."

CERN's LHC has already had considerable success six years ago discovering the existence of the Higgs boson or "God particle" that is believed to endow all other particles with mass. It will continue running experiments in high-energy physics for at least another two decades. The upgrade, which will be known as the High Luminosity LHC, will start collecting data from 2026.

Before then he will have retired, but in the next three years the computing department will continue to grapple with the challenges thrown up by the upgrade with IT researchers exploring innovative computing technologies such as quantum computing to help deal with its future challenges.

The scale of the IT department's task in terms of safely storing and analysing the data from the experiments cannot be underestimated. He predicts CERN will be storing 10s of exabytes (an exabyte is 1bn gigabytes) of data by the time the LHC stops running in 20 or so years' time.

It's an enormous task, but I doubt he will ever lose his passion for the science and being part of the team that helps enable the physics and the research to happen.

"I find physics totally fascinating, while other things I just find boring," he added.

His advice for those hoping to enter the field of scientific research was to stay focused and hard working at the beginning of your career and be willing to learn from others with more experience.

"Now I'm in the position to give things back and that can make a huge difference to a career. Without education things can go badly wrong," he said. "If you don't make the effort in your early career then it can be hard to catch up later." He added that it was important to stand out from the crowd of hopefuls when applying for jobs or being interviewed and don't be afraid of hard work. "Even people like the great Roger Federer works pretty hard," he added.

Preservation & Archiving

- A widely used reference model (ISO 14721) exists for data preservation and archiving
  - **Preservation: intent to re-use; Archiving: hope not!**
- **Also certification procedures (e.g. ISO 16363) based on same standard**
- CERN is pursuing certification for all its LTDP activities: most of the metrics concern the host (lab) of the "trustworthy digital repository"
- **Goal: have LTDP written into the fabric of the organisation as part of 2019/20 ESPP**

F.A.I.R. Data Management

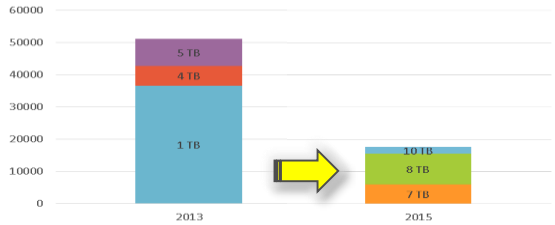
- Increasing emphasis on FAIR DMPs, including preservation, sharing, reproducibility etc.
  - **FAIR now includes also s/w but not yet build systems, verification procedures & environment**
- IMHO not yet fully understood (some claim otherwise) - we see (ir)regular changes on how we **find** data and 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)**

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

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}

Software Preservation

- HEP has since long shared its software across international collaborations
  - **CERNLIB – first started in 1964 and used by many communities worldwide**
- Today HEP s/w is O(10^7) lines of code, 10s to 100s of modules and many languages! (**No standard app**)
- ➢ **Versioning filesystems and virtualisation** look promising: have demonstrated resurrecting s/w 15 years after data taking and hope to provide stability 5-15 years into the future
- ➢ Believe we can analyse LEP data **~30 years** after data taking ended! (i.e. until ~2030)
- **Does anyone have a better idea?**

What is HEP data?

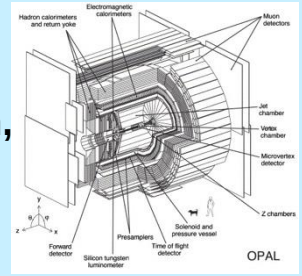
Digital information

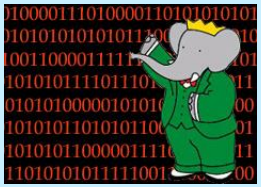
The data themselves, volume estimates for preservation data of the order of a few to 10 PB (to 10 EB for LHC)

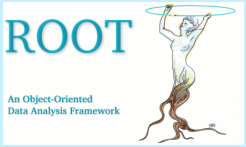
Other digital sources such as databases to also be considered

Software

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





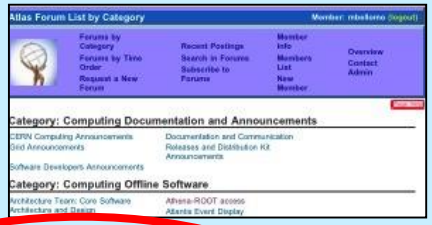


CERNLIB Access

- Access to the CERN Program Library is free of charge to all HEP users worldwide.
- Non-HEP academic and not-for-profit organizations: 1KSF/year

Meta information

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



Publications

arXiv.org

INSPIRE

Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

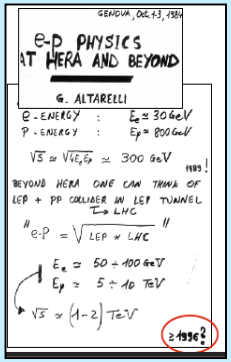
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,)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

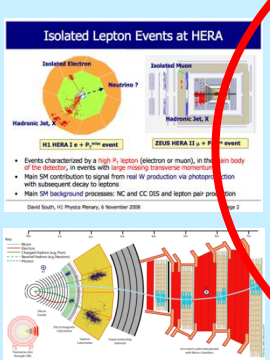
Journal of High Energy Physics

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

Documentation

Internal publications, notes, manuals, slides





Expertise and people

DESY

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 technical knowledge in addition to reprocessing LaTeX into PDF!**

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)**

➢ **"Stage 1" external audit to high-light those areas requiring attention has now been performed**
  - **One particularly weak area is the lack of a formal LTDP policy for Scientific Data (exists in other areas…)**

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)**

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 – no new analyses included

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**

- Getting the degree of public recognition as at the Higgs discovery day was a **target 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 (Science Europe)

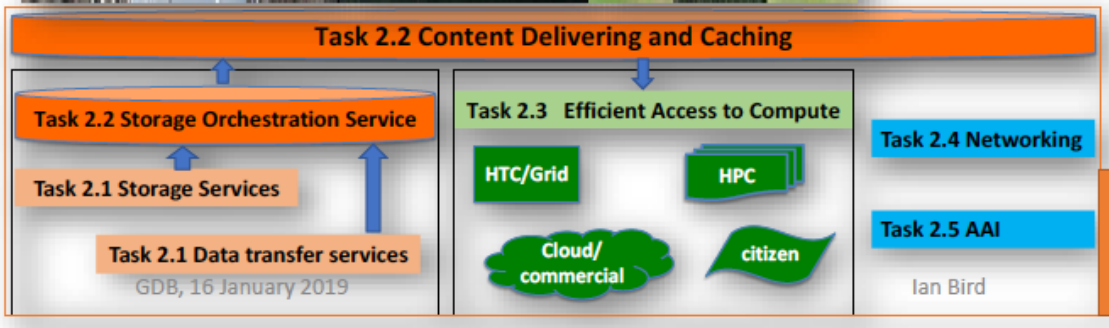
An H2020 project aimed at clusters of ESFRIs

ESFRI Science Projects

HL-LHC	SKA
FAIR	CTA
KM3Net	JIVE-ERIC
ELT	EST
EURO-VO	EGO-VIRGO
(LSST)	(CERN,ESO)

ESCAPE

European Science Cluster of Astronomy & Particle physics
ESFRI research infrastructures



Goals:

Prototype an infrastructure for the EOSC that is adapted to the Exabyte-scale needs of the large ESFRI science projects.

Ensure that the science communities drive the development of the EOSC.

Has to address *FAIR* data management, long term preservation, open access, open science, and contribute to the EOSC catalogue of services.

Work Packages

WP2 – Data Infrastructure for Open Science
WP3 – Open-source scientific Software and Service Repository
WP4 – Connecting ESFRI projects to EOSC through VO framework
WP5 – ESFRI Science Analysis Platform

Data centres (funded in WP2)
CERN, INFN, DESY, GSI, Nikhef, SURFSara, RUG, CCIN2P3, PIC, LAPP, INAF

Objectives

Problem to Solve	**Moving from basic "archiving" to higher level functionality: Long-term DP meets FAIR principles**
What' s missing	**Proven functionality** for **PB range** scientific archives; **Long-term** data services, migration, portability… **(GB/s ingest rates)**
Required R&D	Provide missing functionality preferably using commons solutions and not bespoke
Market potential	Commercial solutions exist at a modest range scale (Preservica, Arkivum, etc.) Communities (DuraSpace, DCC/DPC, DANS) brought together by several conferences

OAIS has a useful definition of "Long Term"

Use Cases

- Archiving 2PB of data from BaBar@SLAC
  - **To be ejected from SLAC site!**
- The DPHEP Science Demonstrator Use Case from the EOSC Pilot
- Additional Use Cases from “photon science”, life science, astronomy, ESA and others
- Clearly not aiming to provide a solution for millennia but could be of immediate interest to e.g. Museum Archives
  - **Dinosaurs, Egyptian artifacts etc.**

Suggestions

- An invited talk at PV2020 at CERN?
- Consider joining the Digital Preservation Coalition?
- Use Case(s) for ARCHIVER?

Digital Preservation Coalition

- CERN recently joined, mainly in the context of the ARCHIVER project
- Additional benefits have already been seen (that were not completely obvious prior to joining)…
- (Surprising) synergies with widely differing domains!
- Will host iPRES 2022

Open consultation

Nuclear Decommissioning Authority: Business Plan 2019 to 2022

Published 3 December 2018
From: Nuclear Decommissioning Authority, Radioactive Waste Management, Sellafield Ltd, Dounreay , Magnox Ltd, and Low Level Waste Repository Ltd

Summary

The Business Plan sets out key activities and expected progress for all 17 of the NDA's nuclear sites over the next 3 years.

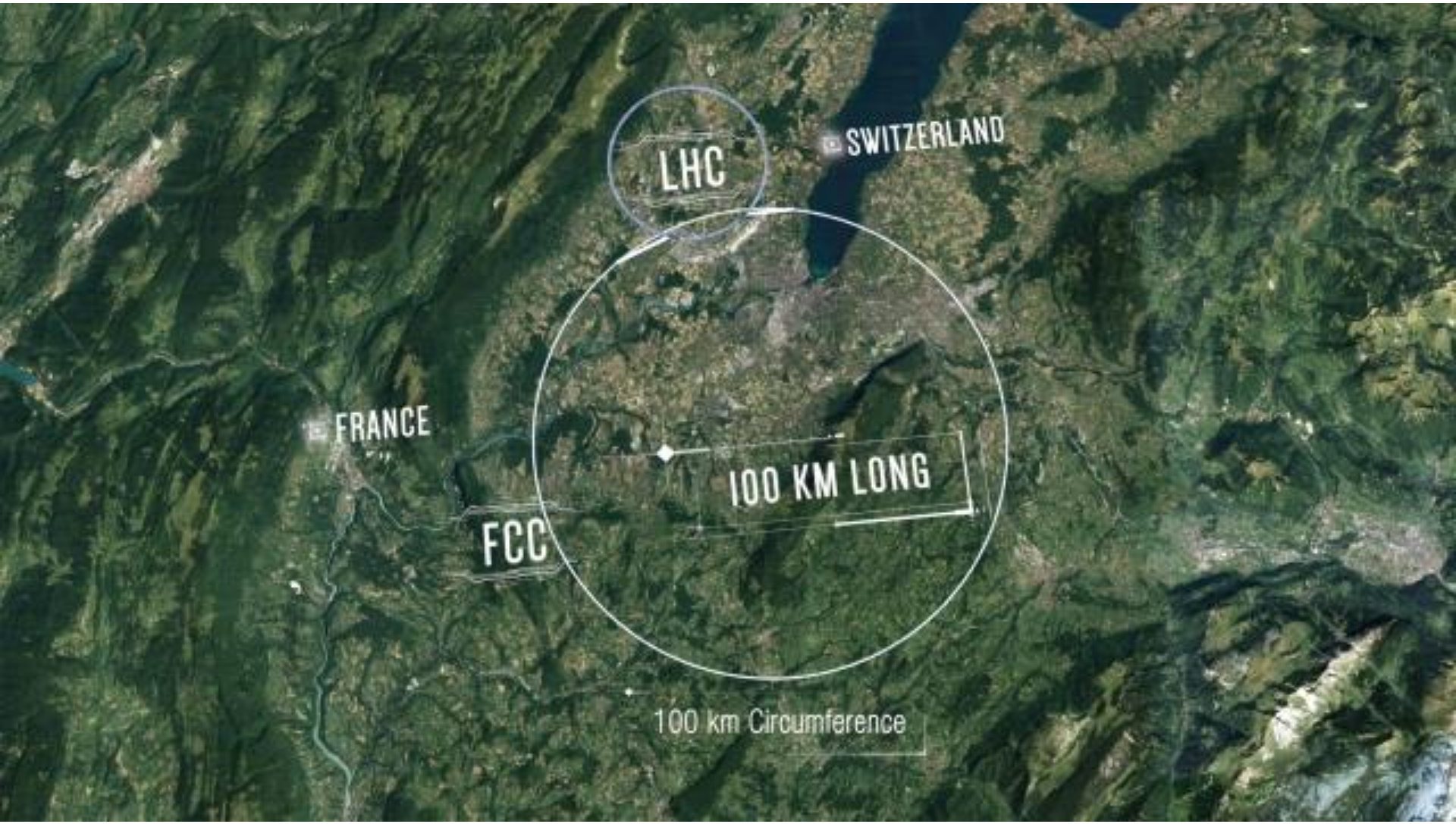
This consultation closes at
4pm on 4 February 2019

Summary

- There are a number of potential services (and providers) to assist with LTDP
- But they are not "holistic" - "mind the gap(s)"
- **Information is >> data**
- There are many fora to join!
  - e.g. DPC, iPRES, PASIG, PV, iDCC, ...

- Beware of people offering a perfect solution
  - Particularly those in IT, who love change!
- Beware of "grey-backed gorillas"

==> Constant effort is needed – and long-term commitment


LHC
SWITZERLAND
FRANCE
FCC
100 KM LONG
100 km Circumference