



Long-Term Data Preservation in HEP

Challenges, Opportunities and Solutions(?)

Jamie.Shiers@cern.ch

Workshop on Best Practices for Data
Management & Sharing



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

Overview

- What we are trying to do – and why
- How we plan to address the problem(s)
- Where we have “solutions” or “expertise” that might help others
- Collaboration, funding & sustainability

Collaboration

- High Energy Physics (HEP) is build around **collaboration** and **collaborations**
- Even after some 30+ years, it is still at times incredible to me how institutes and people from all around the world can work together to build things as complex as HEP accelerators and detectors – and produce results!
- This is challenging enough today – but think of the era when “communication” was done by Telex!

Collaboration Before the Web!





ATLAS
Collaboration

Argentina	Morocco
Armenia	Netherlands
Australia	Norway
Austria	Poland
Azerbaijan	Portugal
Belarus	Romania
Brazil	Russia
Canada	Serbia
Chile	Slovakia
China	Slovenia
Colombia	South Africa
Czech Republic	Spain
Denmark	Sweden
France	Switzerland
Georgia	Taiwan
Germany	Turkey
Greece	UK
Israel	USA
Italy	CERN
Japan	JINR



July 2010

ATLAS consists of ~2900 persons from 37 countries, more than 170 institutes / universities



KEK

Fermilab

FRAGILE

tabo

- 60 years of Science for Peace
- Scientific Discoveries and the way ahead
- International Cooperation –
Science as a motor for international dialogue
- 60 years of progress in Science and Technology
- Science Education and Training –
modern science for everyone

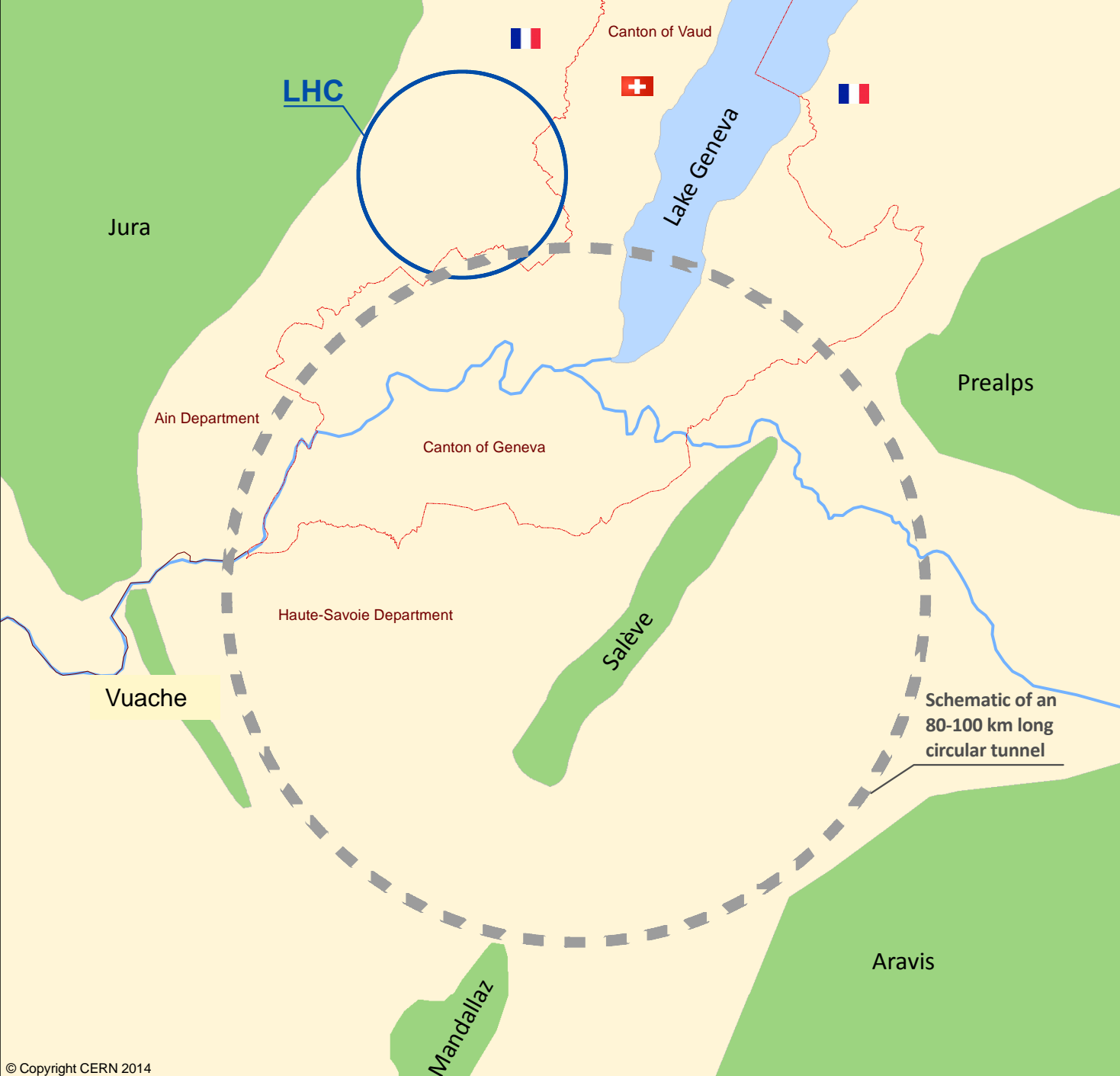


Collaboration – Summary

- Highly successful international collaboration(s) over **more than 6 decades** – often under difficult circumstances
 - This includes not only **intra-HEP** work but also that with **other disciplines** – as well as the application of technologies to other areas
- **We have benefited a lot from collaboration!**

Funding & Sustainability

- **“The success of the LHC is proof of the effectiveness of the European organisational model for particle physics, founded on the sustained long-term commitment of the CERN Member States and of the national institutes, laboratories and universities closely collaborating with CERN.**
- ***Europe should preserve this model in order to keep its leading role, sustaining the success of particle physics and the benefits it brings to the wider society.”***
- European Strategy for Particle Physics – 1st update
- Adopted at 16th European Strategy Session of Council
- Brussels, May 2013
- <https://indico.cern.ch/event/244974/page/1>



Balance sheet

- 20 year investment in Tevatron ~ \$4B
- Students \$4B
- Magnets and MRI \$5-10B } ~ \$50B total
- Computing \$40B

Very rough calculation – but confirms our gut feeling that investment in fundamental science pays off

I think there is an opportunity for someone to repeat this exercise more rigorously

cf. STFC study of SRS Impact

<http://www.stfc.ac.uk/2428.aspx>



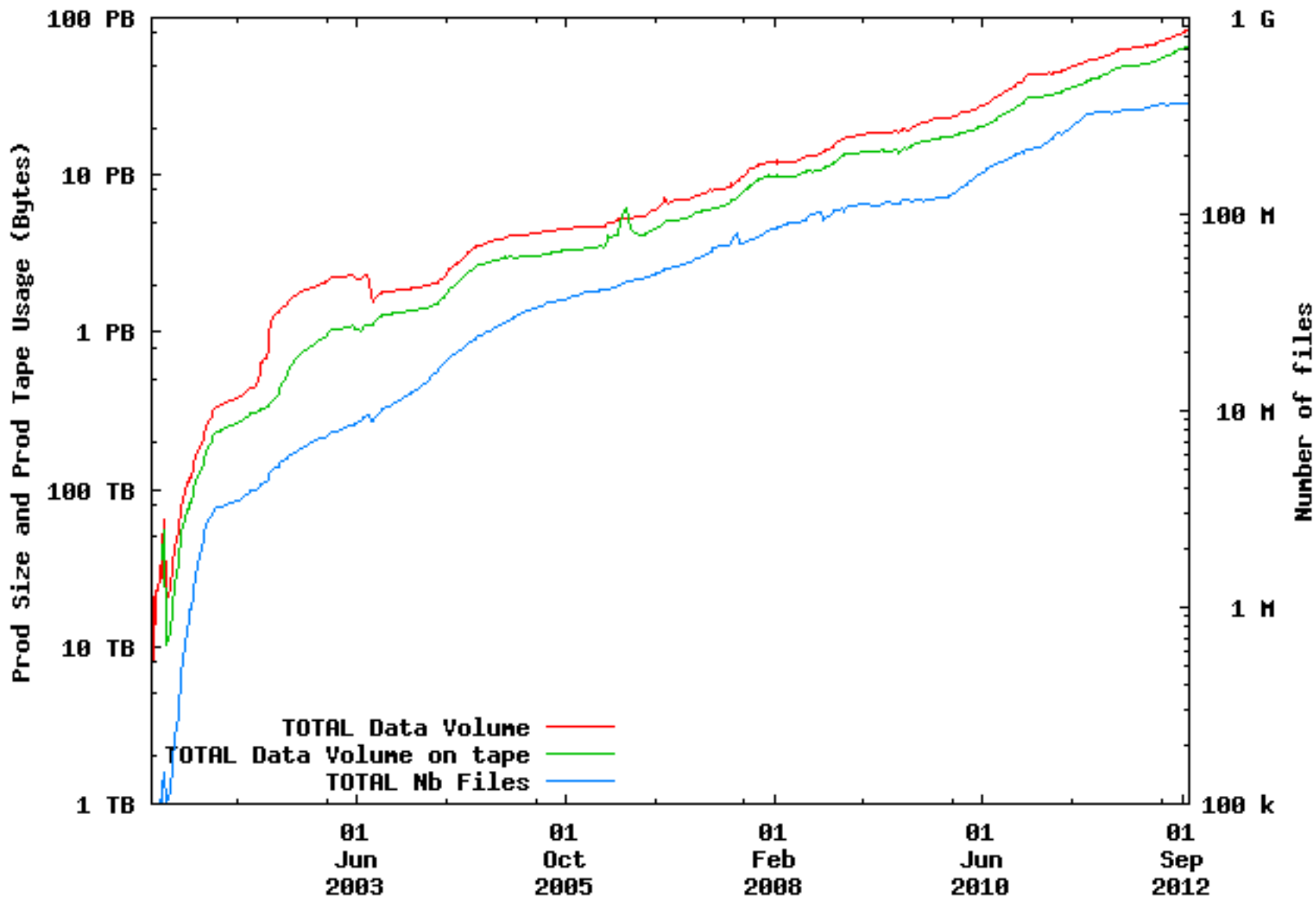
Funding & Sustainability



- **We rely on public money**
- We have a close relationship with the funding agencies, who are tightly involved in the approval of the scientific programme and the details of how resources are allocated
- [And they are requiring us to “do data preservation” – and show them the results!]
- We have not only a **strategy** but also a **strategy** for updating the **strategy**!
- [Long may this continue]

CERN has ~100 PB archive

Experiments Production Data in CASTOR

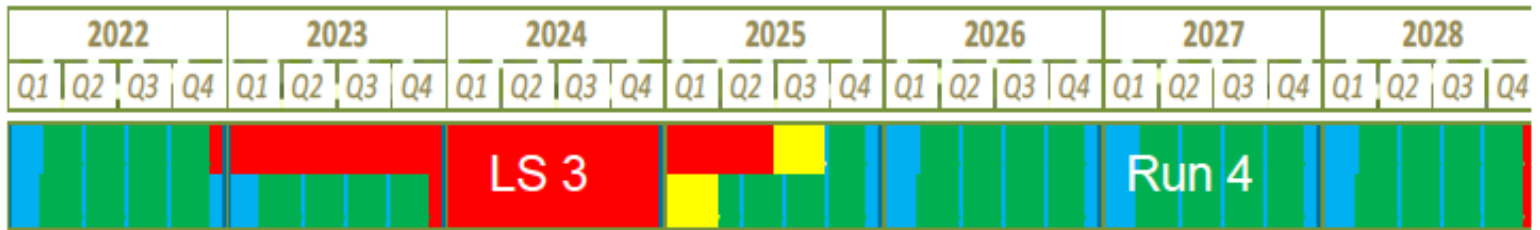
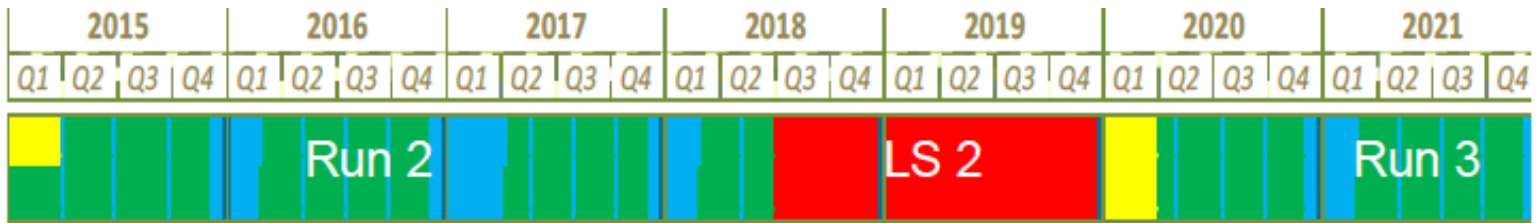


Generated Sep 25, 2012 CASTOR (c) CERN/IT

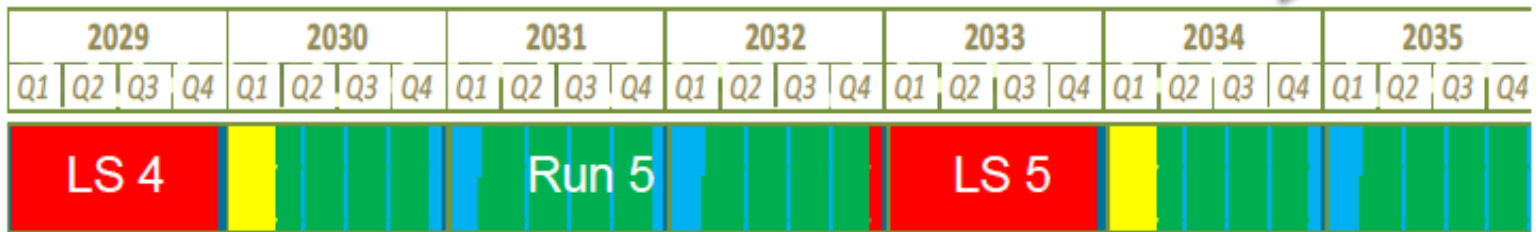
LHC schedule beyond LS1

Run 1 – which led to the discovery of the Higgs boson – is just the beginning. There will be further data taking – possibly for another 2 decades or more – at increasing data rates, with further possibilities for discovery!

We are here

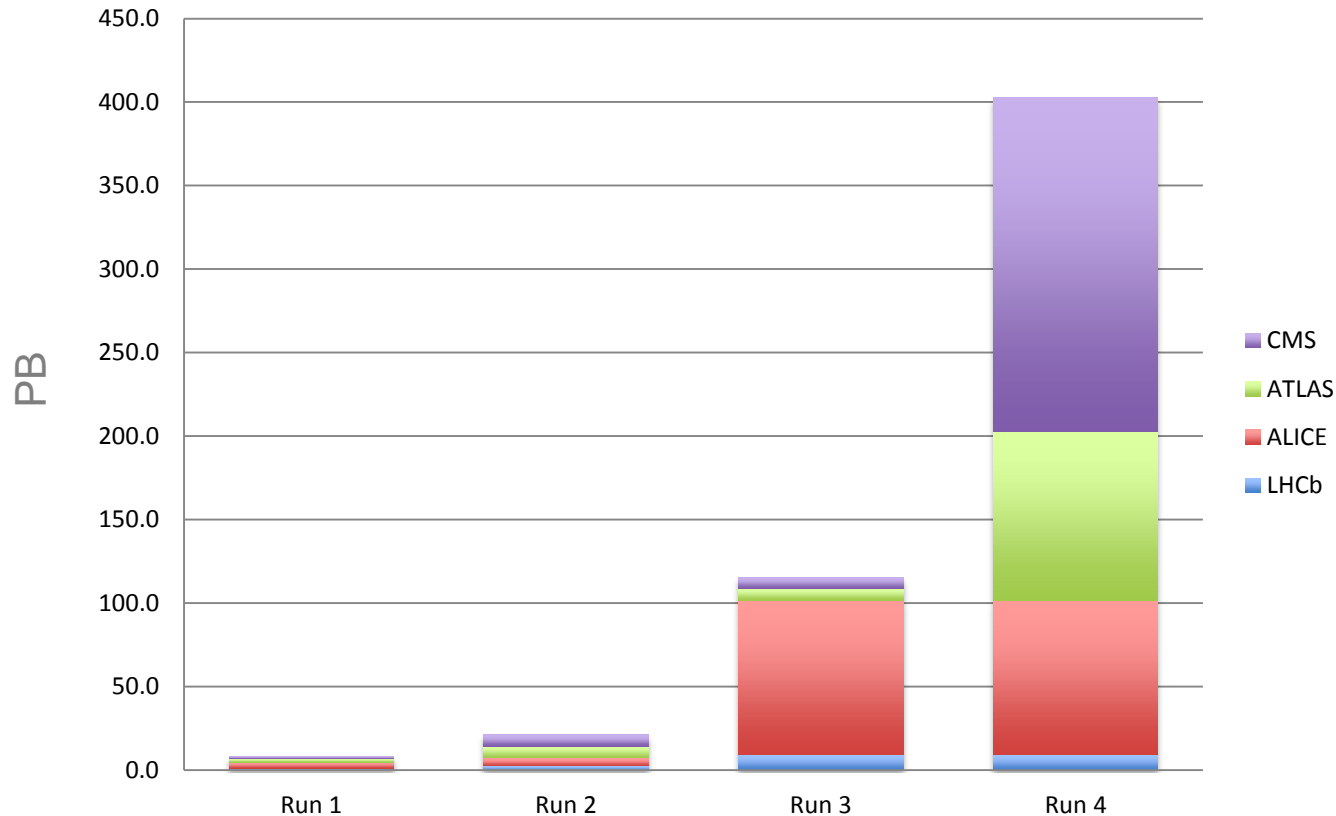


→ HL-LHC





Data: Outlook for HL-LHC

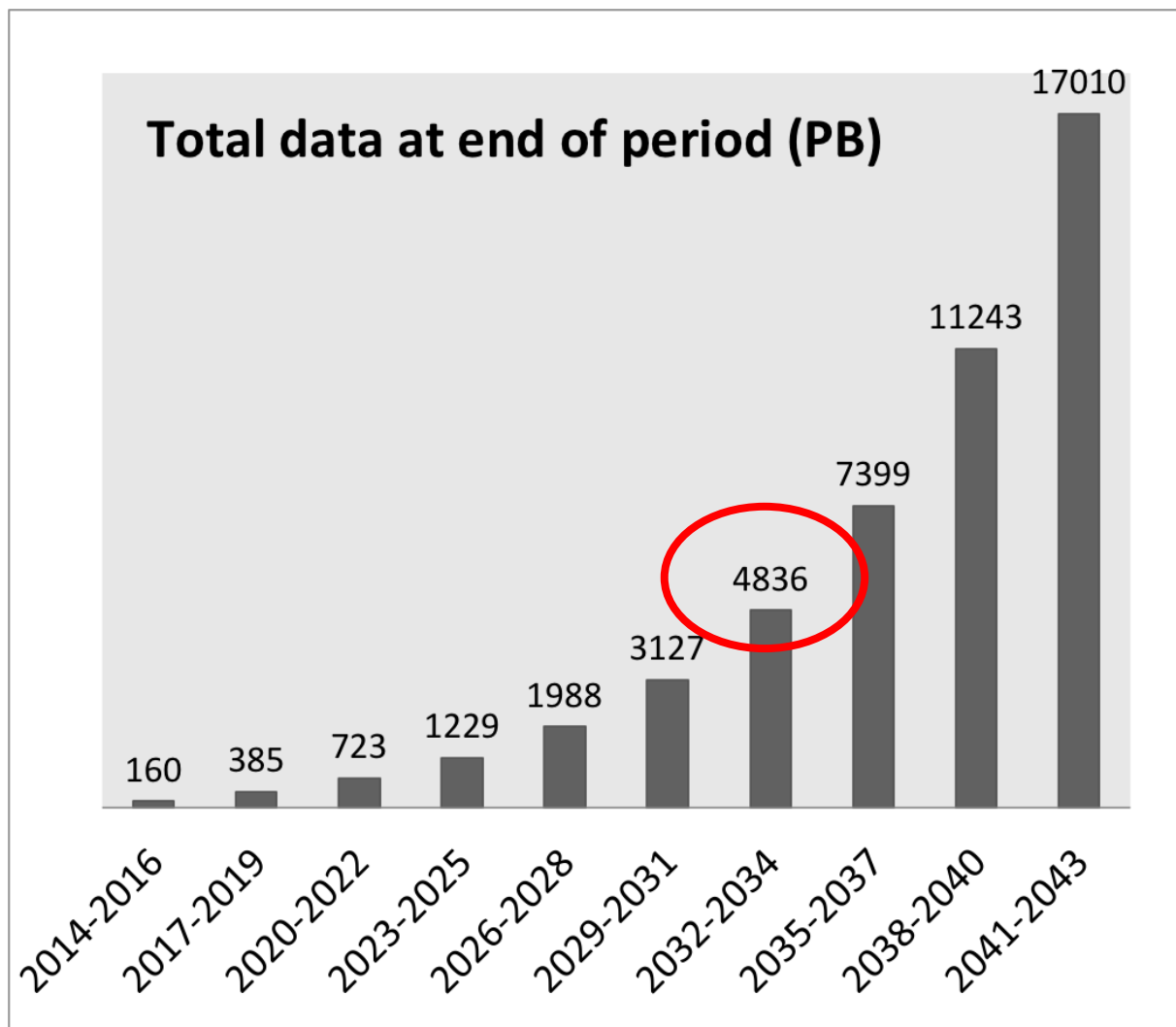


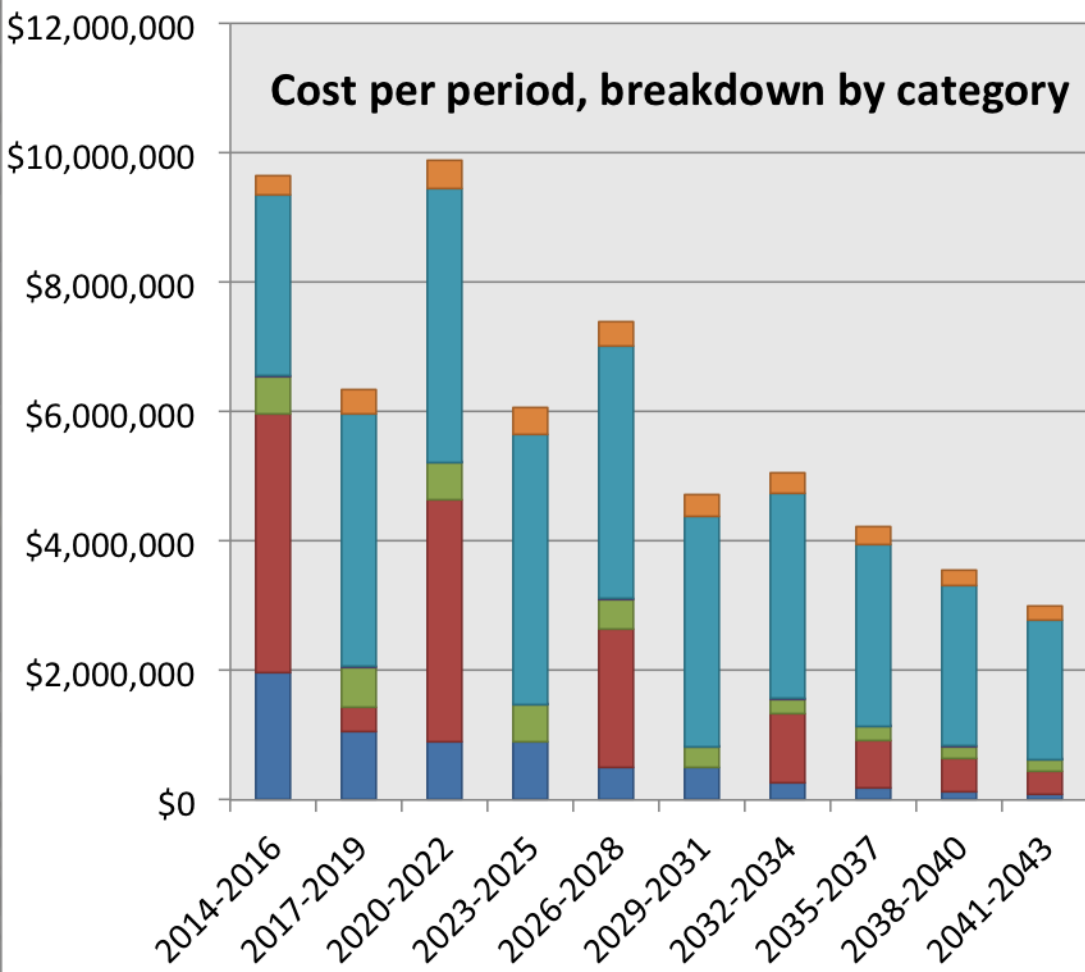
- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
 - To be added: derived data (ESD, AOD), simulation, user data...
- **0.5 EB / year is probably an under estimate!**



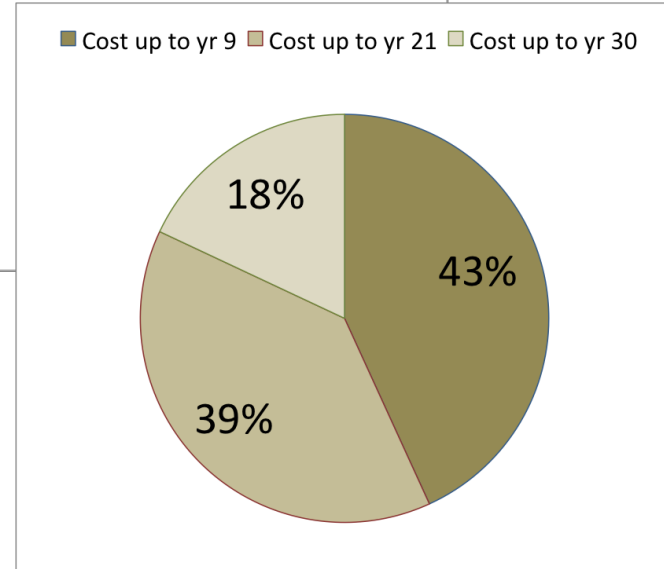
DSS Cost Modelling: Regular Media Refresh + Growth

Start with 10PB, then +50PB/year, then +50% every 3y (or +15% / year)

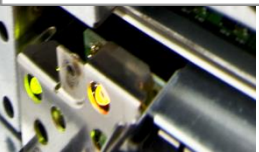




- Total period disk server power cost
- Total period disk server hardware+maint cost
- Total period tape power cost
- Total period tape maintenance cost
- Total period tape media cost
- Total period tape hardware cost



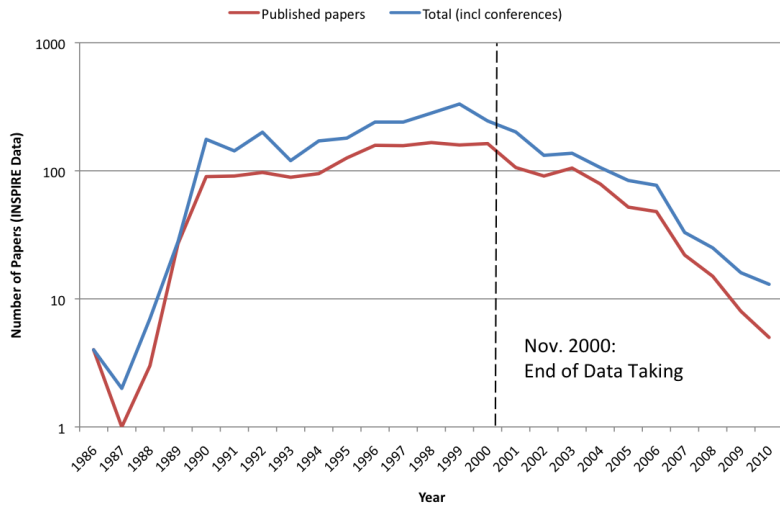
**Total cost: ~\$60M
 (~\$2M / year)**



2020 Vision for LT DP in HEP

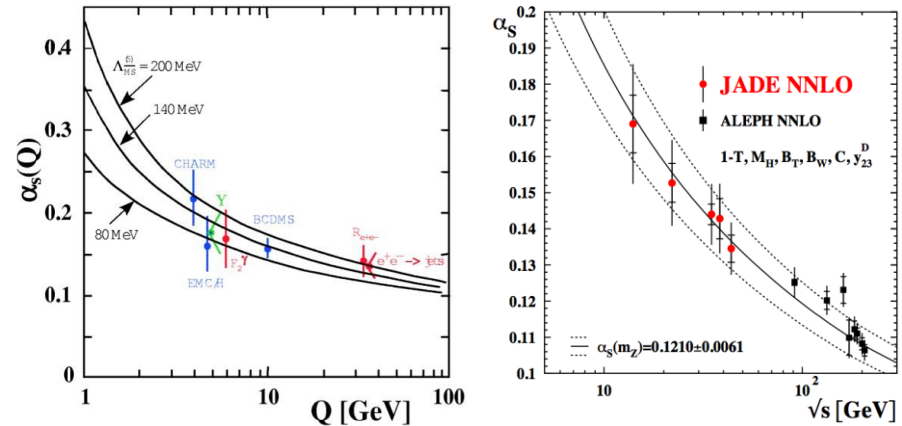
- Long-term – e.g. LC timescales: disruptive change
 - By 2020, all archived data – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to annotate further
 - Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards
 - **DPHEP portal**, through which data / tools accessed
- **Agree with Funding Agencies clear targets & metrics**

1 - Long Tail of Papers



3

2 - New Theoretical Insights

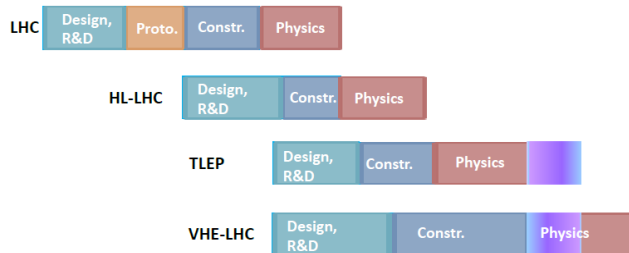


4

3 - "Discovery" to "Precision"



possible long-term time line



5

Use Case Summary

1. Keep data usable for ~1 decade
2. Keep data usable for ~2 decades
3. Keep data usable for ~3 decades

Volume: 100PB + ~50PB/year (+400PB/year from 2020)

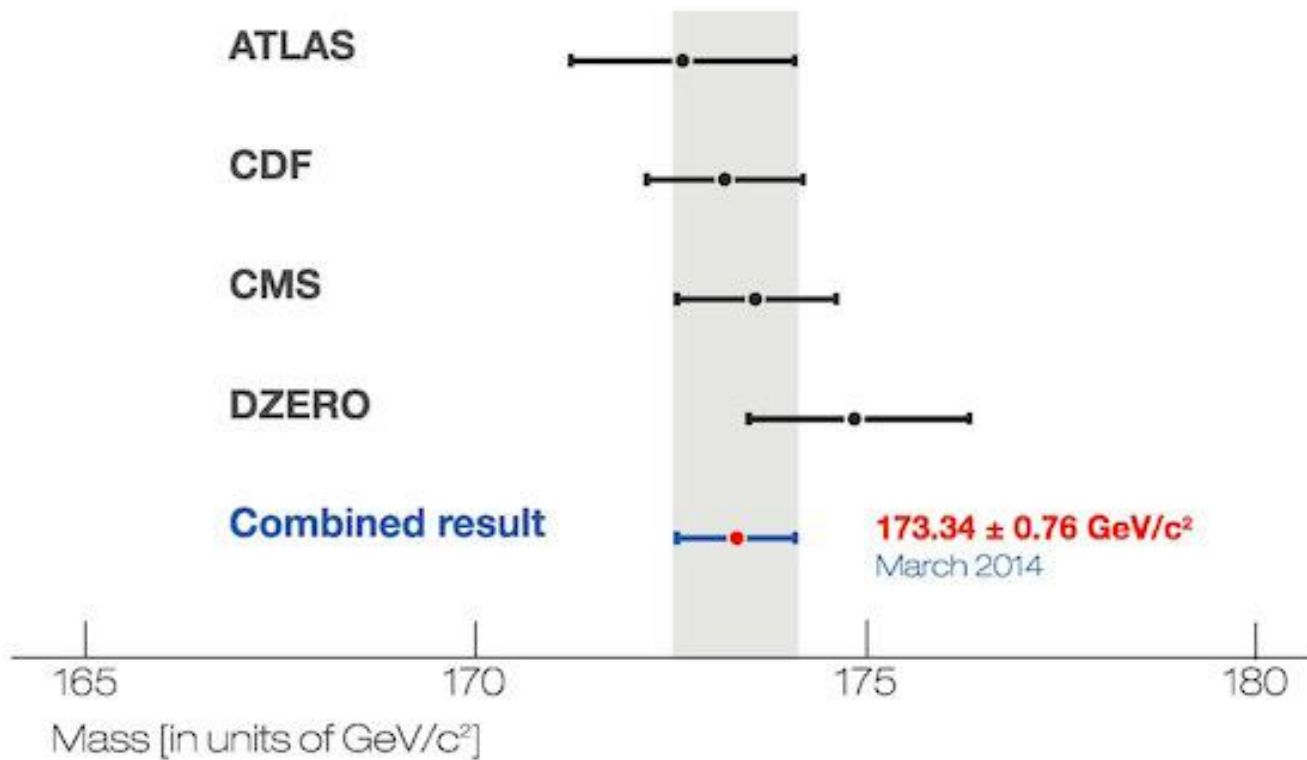
7

Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the Office of Science for research funding are required to include a Data Management Plan (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be **shared and preserved** or explains why data sharing and/or preservation are not possible or scientifically appropriate.
- At a minimum, DMPs must describe how data sharing and preservation will enable **validation of results**, or how results could be validated if data are not shared or preserved.
- Similar requirements from European FAs and EU (H2020)

20 Years of the Top Quark

Top quark mass measurements





How?

- How are we going to preserve all this data?
- And what about “the knowledge” needed to use it?
- How will we measure our success?
- And what’s it all for?

Answer: Two-fold

- Specific technical solutions
 - Main-stream;
 - Sustainable;
 - Standards-based;
 - **COMMON**
- Transparent funding model
- Project funding for short-term issues
 - Must have a plan for long-term support from the outset!
- Clear, up-front metrics
 - Discipline neutral, where possible;
 - Standards-based;
 - **EXTENDED IFF NEEDED**
- Start with “the standard”, coupled with recognised certification processes
 - See RDA IG
- Discuss with FAs and experiments – agree!
- (For sake of argument, let’s assume DSA)

Additional Metrics (aka “The Metrics”)

1. Open Data for educational outreach

- Based on specific samples suitable for this purpose
- **MUST EXPLAIN BENEFIT OF OUR WORK FOR FUTURE FUNDING!**
- High-lighted in European Strategy for PP update

2. Reproducibility of results

- A (scientific) requirement (from FAs)
- **“The Journal of Irreproducible Results”**

3. Maintaining full potential of data for future discovery / (re-)use

LHC Data Access Policies

Level (standard notation)	Access Policy
L0 (raw) (cf “Tier”)	Restricted even internally <ul style="list-style-type: none">• Requires significant resources (grid) to use
L1 (1 st processing)	Large fraction available after “embargo” (validation) period <ul style="list-style-type: none">• Duration: a few years• Fraction: 30 / 50 / 100%
L2 (analysis level)	Specific (meaningful) samples for educational outreach: pilot projects on-going <ul style="list-style-type: none">• CMS, LHCb, ATLAS, ALICE
L3 (publications)	Open Access (CERN policy)

1. DPHEP Portal


2. **Digital library** tools (**Invenio**) & services (**CDS, INSPIRE, ZENODO**) + domain tools (**HepData, RIVET, RECAST...**)
3. **Sustainable software**, coupled with advanced **virtualization** techniques, “snap-shotting” and **validation** frameworks
4. **Proven bit preservation** at the 100PB scale, together with a **sustainable** funding model with an outlook to 2040/50
(and several EB of data)
5. **Open Data**

DPHEP Portal – Zenodo like?

The screenshot shows the Zenodo website interface. At the top, there is a search bar with a magnifying glass icon and a 'Search' button. Below the search bar, there is a 'Filter by types' section. The main content area is titled 'Recent Uploads' and lists three items:

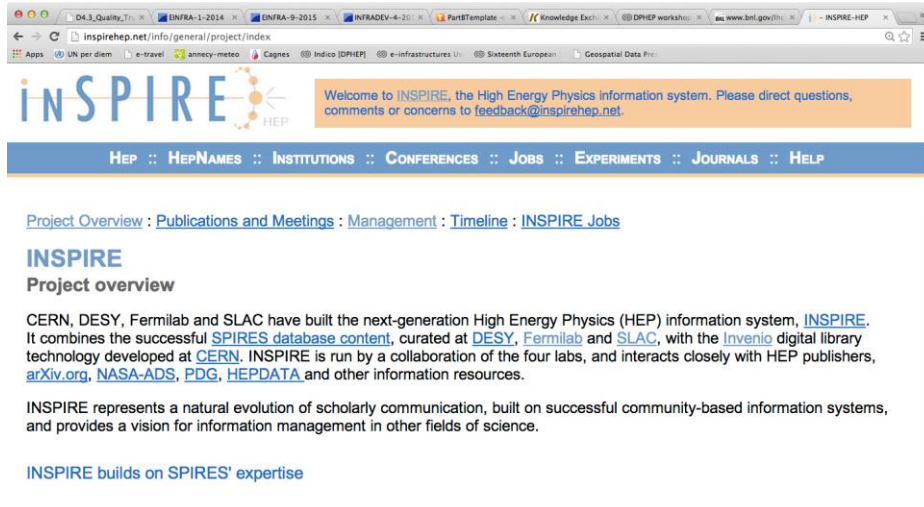
- 01 April 2014** **Journal article** **Open access** [View](#)
The e-book phenomenon: a disruptive technology
Tom D. Wilson
The emergence of the e-book as a major phenomenon in the publishing industry is of interest, world-wide. The English language market, with Amazon.com as the major player in the market may have dominated attention, but the e-book has implications for many ...
- 10 November 2010** **Thesis** **Open access** [View](#)
Χρόνιες Επιδράσεις του Καπνίσματος στη Λειτουργική Ικανότητα του Κυκλοφορικού Συστήματος Νεαρών Υγιών Ατόμων
Papathansiou, George ; Evangelou, Angelos
Εισαγωγή Το κάπνισμα αποτελεί τον σοβαρότερο (ίσως παράγοντα κινδύνου μελλοντικής καρδιοαγγειακής νοσηρότητας και θνητότητας ενώ θεωρείται ως η κυριότερη αντιστρεπτή αιτία θανάτου. Το κάπνισμα συνδέεται με χρονότροπη καθυστέρηση λόγω δυσλειτουργίας του ...
Uploaded by [George](#) on 30 March 2014.
- 30 March 2014** **Report** **Open access** [View](#)
Archaeobotanical remains from Mitchelstown and Ballnamona

On the right side of the page, there are two informational boxes:

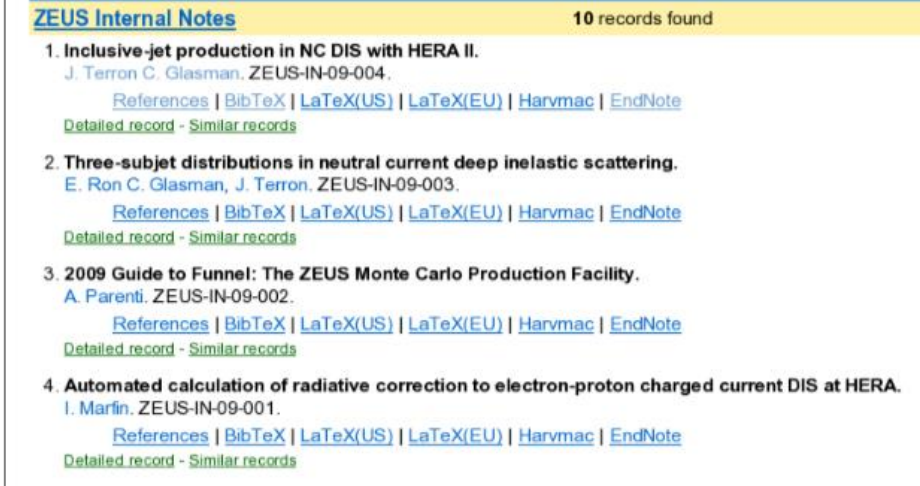
- GitHub integration** 
Want to preview the public beta of GitHub integration? Just [Sign In](#) with your GitHub account and [click here](#).
- New to ZENODO?**
 - **Research. Shared.** – all research outputs from across all fields of science are welcome!
 - **Citeable. Discoverable.** – uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
 - **Community Collections** – accept or reject uploads to your own community collections (e.g workshops, EU projects or your complete own digital repository).
 - **Funding** – integrated in reporting lines for research funded by the European Commission via OpenAIRE.
 - **Flexible licensing** – because not everything is under Creative Commons.
 - **Safe** – your research output is stored safely for the future in same cloud infrastructure as research data from CERN's Large Hadron Collider.
 - **DropBox integration** – upload files straight from your DropBox.

Documentation projects with INSPIREHEP.net

- Internal notes from all HERA experiments now available on INSPIRE
 - A collaborative effort to provide “consistent” documentation across all HEP experiments – starting with those at CERN – as from 2015
 - (Often done in an inconsistent and/or ad-hoc way, particularly for older experiments)



The screenshot shows the INSPIRE website interface. At the top, there is a navigation bar with links for HEP, HEP NAMES, INSTITUTIONS, CONFERENCES, JOBS, EXPERIMENTS, JOURNALS, and HELP. Below this, a section titled "INSPIRE Project overview" provides a brief description of the system, mentioning its development by CERN, DESY, Fermilab, and SLAC. It also lists various information resources like SPIRES, arXiv, and HEPDATA. A link to "INSPIRE builds on SPIRES' expertise" is also visible.



The screenshot displays a search result for "ZEUS Internal Notes" with 10 records found. The results are listed in a numbered format, each with a title, author, and a set of links for references and detailed records. The records are:

- 1. Inclusive-jet production in NC DIS with HERA II.**
J. Terron C. Glasman. ZEUS-IN-09-004.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)
- 2. Three-subjet distributions in neutral current deep inelastic scattering.**
E. Ron C. Glasman, J. Terron. ZEUS-IN-09-003.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)
- 3. 2009 Guide to Funnel: The ZEUS Monte Carlo Production Facility.**
A. Parenti. ZEUS-IN-09-002.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)
- 4. Automated calculation of radiative correction to electron-proton charged current DIS at HERA.**
I. Marfin. ZEUS-IN-09-001.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)





The Guidelines 2014-2015

Guidelines Relating to Data Producers:

1. The data producer deposits the data in a data repository with **sufficient information** for others to assess the **quality** of the data and compliance with disciplinary and ethical norms.
- 2. The data producer provides the data in formats recommended by the data repository.**
3. The data producer provides the data together with the **metadata** requested by the data repository.



Some HEP data formats

Experiment	Accelerator	Format	Status
ALEPH	LEP	BOS	?
DELPHI	LEP	Zebra	CERNLIB – no longer formally supported
L3	LEP	Zebra	“
OPAL	LEP	Zebra	“
ALICE	LHC	ROOT	PH/SFT + FNAL
ATLAS	LHC	ROOT	“
CMS	LHC	ROOT	“
LHCb	LHC	ROOT	“
COMPASS	SPS	Objy	Support dropped at CERN ~10 years ago
COMPASS	SPS	DATE	ALICE online format – 300TB migrated

Other formats used at other labs – many previous formats no longer supported!

Work in Progress...

- By September, CMS should have made a public release of some data + complete environment
 - LHCb, now also ATLAS , plan something similar, based on “common tools / framework”
- By end 2014, a first version of the “DPHEP portal” should be up and running
- “DSA++” – by end 2015???
- More news in Sep (RDA-4) / Oct (APA)

Mapping DP to H2020

- EINFRA-1-2012 “Big Research Data”
 - Trusted / certified federated digital repositories with sustainable funding models that scale from many TB to a few EB
- “Digital library calls”: front-office tools
 - Portals, digital libraries *per se* etc.
- VRE calls: complementary proposal(s)
 - INFRADEV-4
 - EINFRA-1/9

The Bottom Line

- ✓ **We have particular skills in the area of large-scale digital (“bit”) preservation AND a good (unique?) understanding of the costs**
 - Seeking to further this through RDA WGs and eventual prototyping -> sustainable services through H2020 across “federated stores”
- **There is growing realisation that Open Data is “the best bet” for long-term DP / re-use**
- **We are eager to collaborate further in these and other areas...**

Key Metrics For Data Sharing

- 1. (Some) Open Data for educational outreach**
 - 2. Reproducibility of results**
 - 3. Maintaining full potential of data for future discovery / (re-)use**
- “Service provider” and “gateway” metrics still relevant but IMHO secondary to the above!



Data Sharing in Time & Space

Challenges, Opportunities and Solutions(?)

Jamie.Shiers@cern.ch

Workshop on Best Practices for Data
Management & Sharing



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