Long Term Data Preservation

LTDP = Data Sharing – In Time and Space

Jamie.Shiers@cern.ch

Big Data, Open Data Workshop, May 2014

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

Outline

1. The problem(s) that we are trying to address

2. How collaboration is helping us all

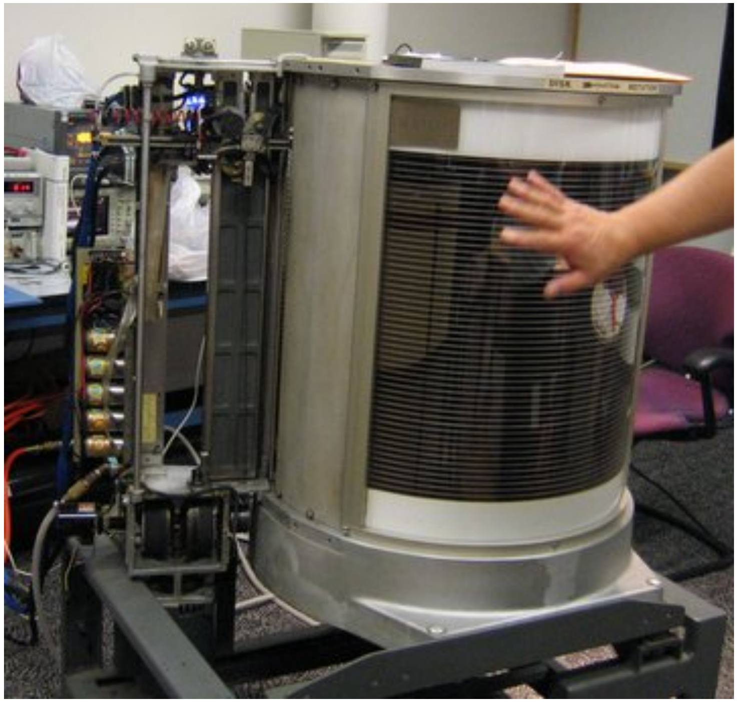
3. Where we could work together to build long-term, scalable and sustainable solutions targetting multi-disciplinary data sharing…

Data Volumes and Outlook

- CERN stores around **100,000 TB** (100PB) of physics data > 2/3 of which is from the LHC
- This will grow to around **5EB** (5,000,000 TB) and beyond in the next 1-2 decades
- Aside from its direct **scientific** potential, it also has significant value for **educational** purposes
- We need to **preserve** it – and the ability to (re-)use it (**share it**), now and in the long-term

➢ This is now both **affordable** as well as **technically** possible

IBM 350 RAMAC



1956, 5 Mch, 8 Kch/s IO

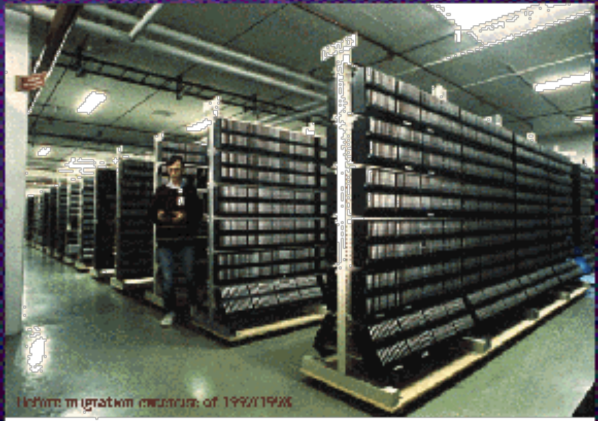
PDP DECtape

1970, 144K 18_ bit words

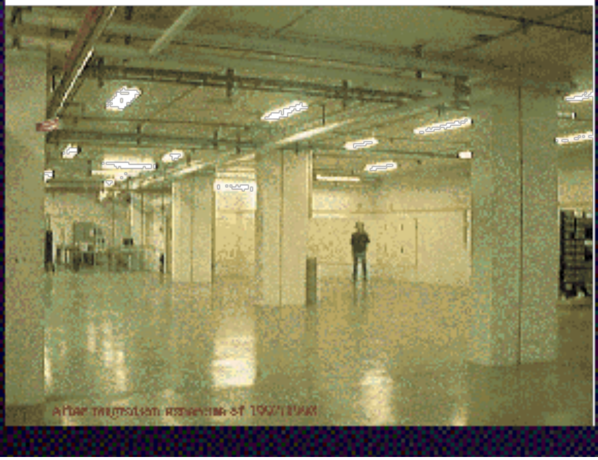
Options

- Ignore problem: we'd like to but….

~300K tapes were 'archived'… .. ~150K were manually mounted….



Before migration [illegible] of 1997/1998

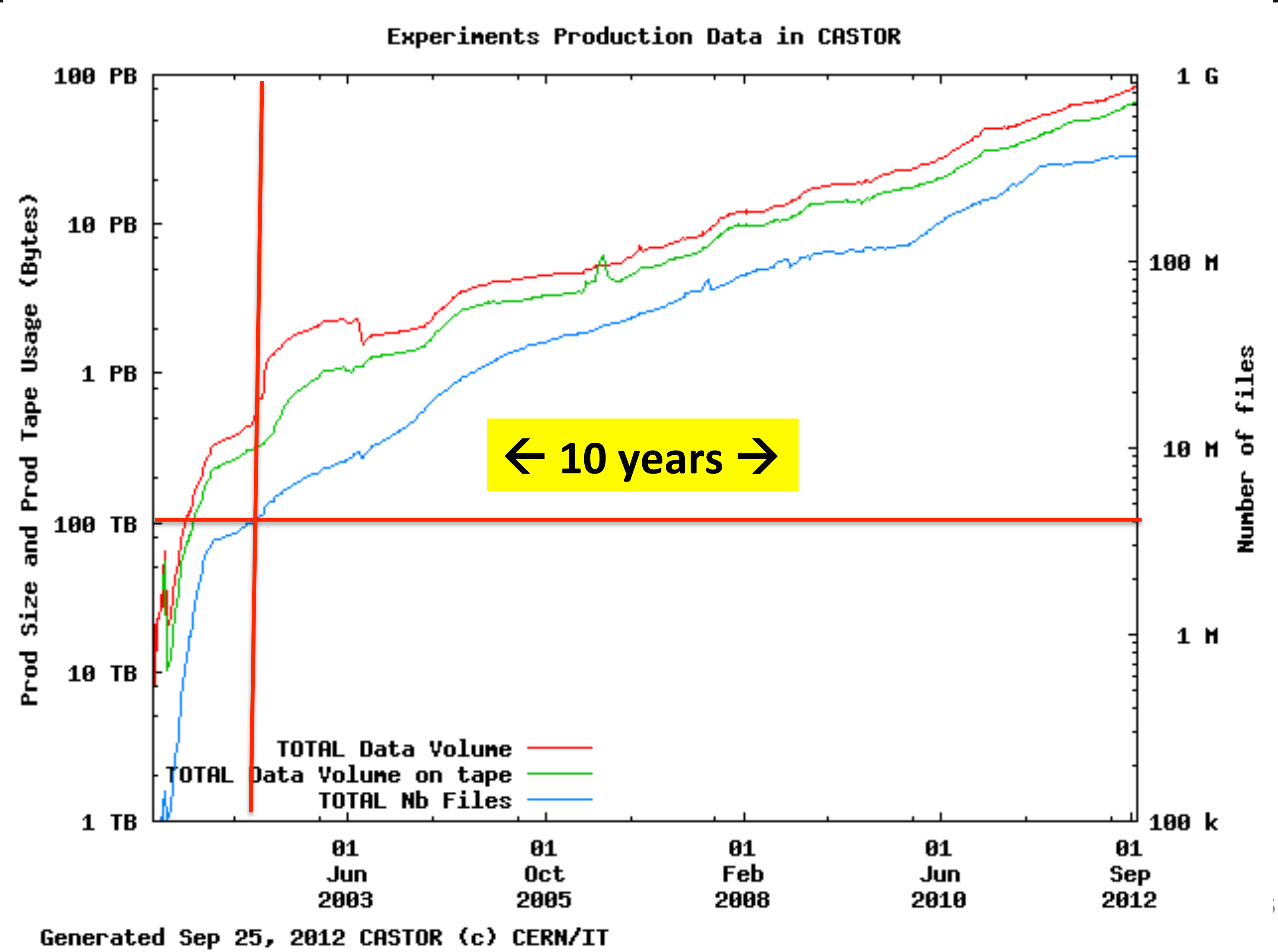


After migration [illegible] of 1997/1998

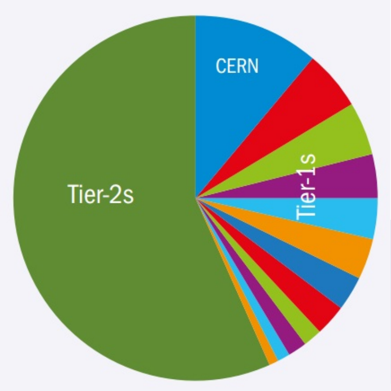
..and then copied to Redwoods….

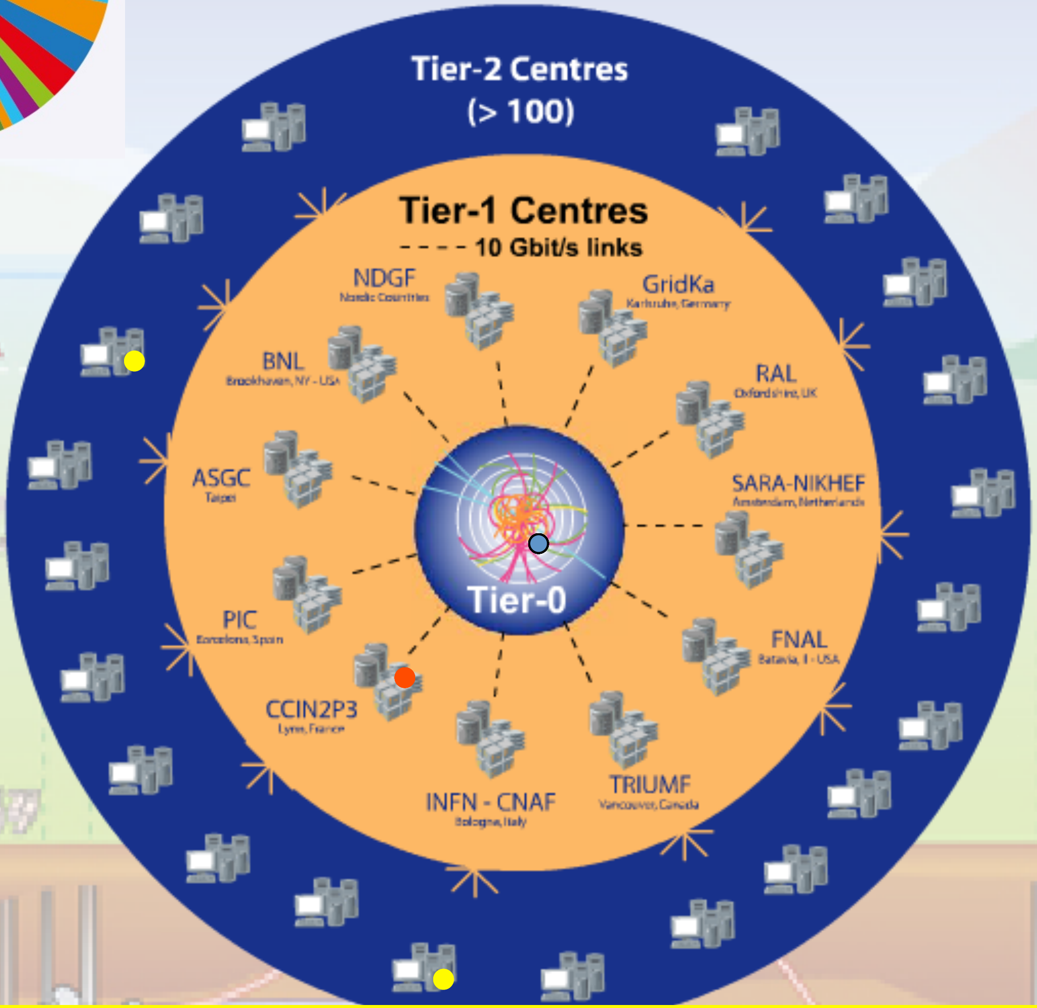
CERN

CERN has ~100 PB archive



WLCG Tier 0 – Tier 1 – Tier 2





Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

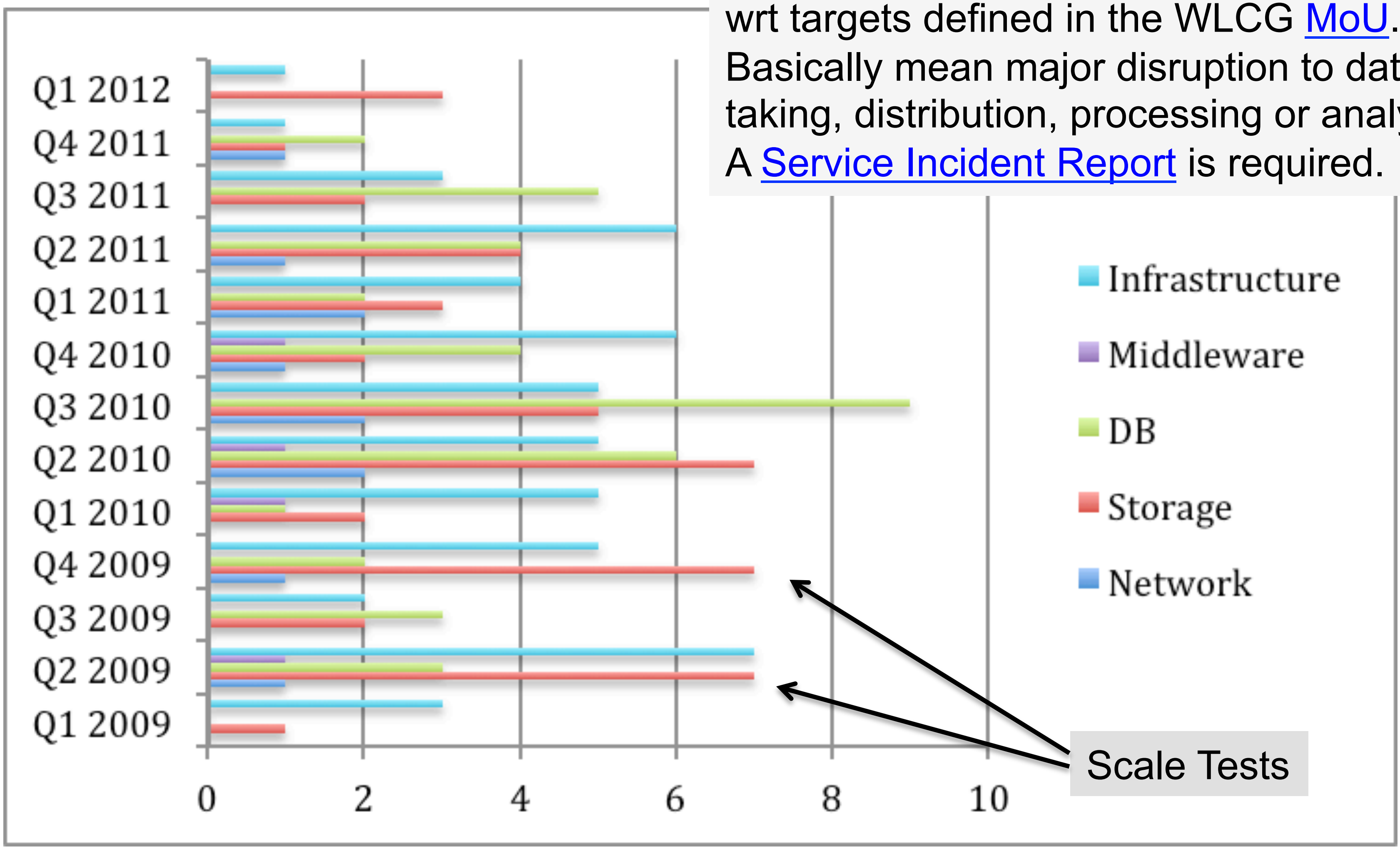
Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis

Extremely effective for data processing and analysis, but not well adapted to LTDP

WLCG Service Incidents

These are significant service incidents wrt targets defined in the WLCG MoU. Basically mean major disruption to data taking, distribution, processing or analysis. A Service Incident Report is required.



Scale Tests

And the grid?

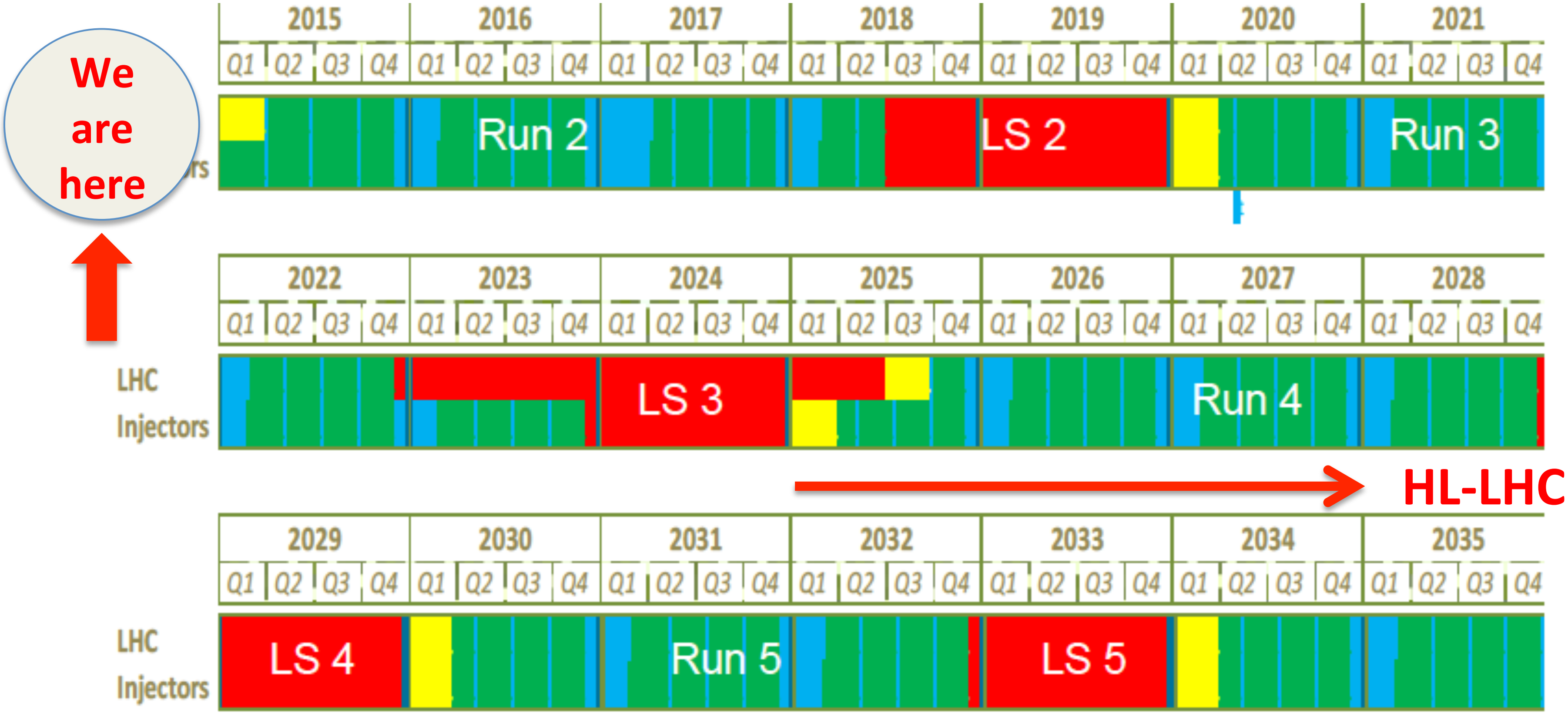
- To find the Higgs you need 3 things:

1. **The machine;**
2. **The experiments;**
3. **The GRID**

- Rolf-Dieter Heuer, DG, CERN, July 4 2012

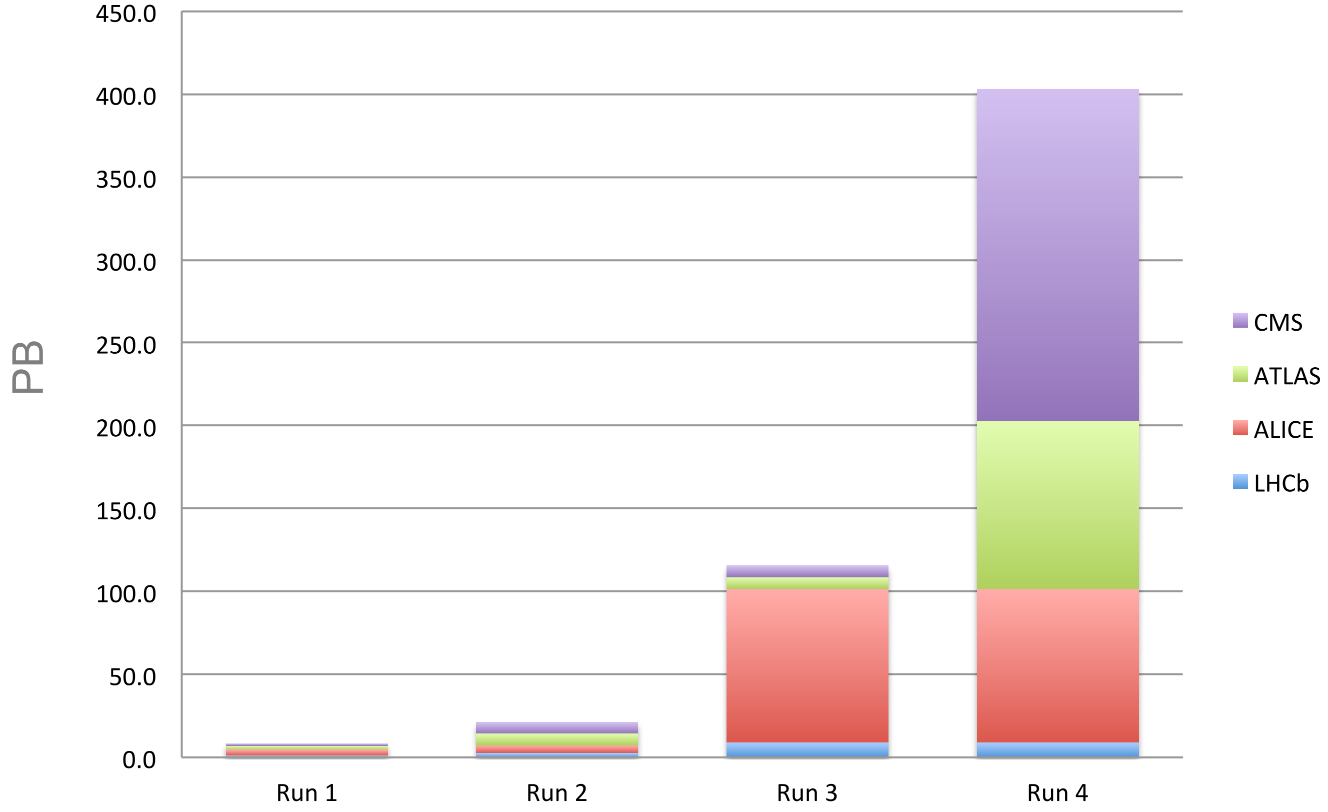
LHC schedule beyond LS1

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!



LHC schedule approved by CERN management and LHC experiments spokespersons and technical coordinators
Monday 2nd December 2013

Data: Outlook for HL-LHC

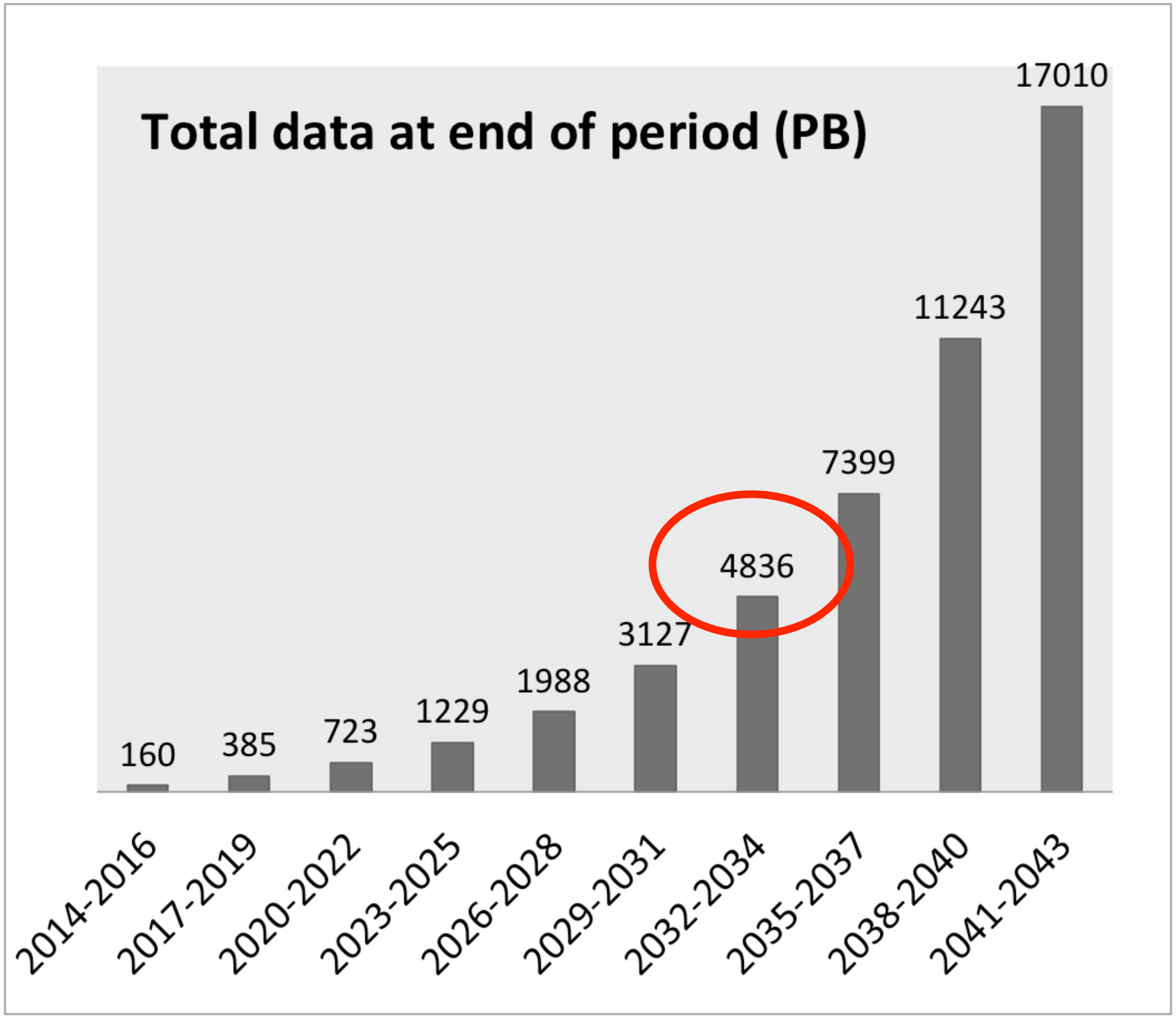


- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
  - To be added: derived data (ESD, AOD), simulation, user data…

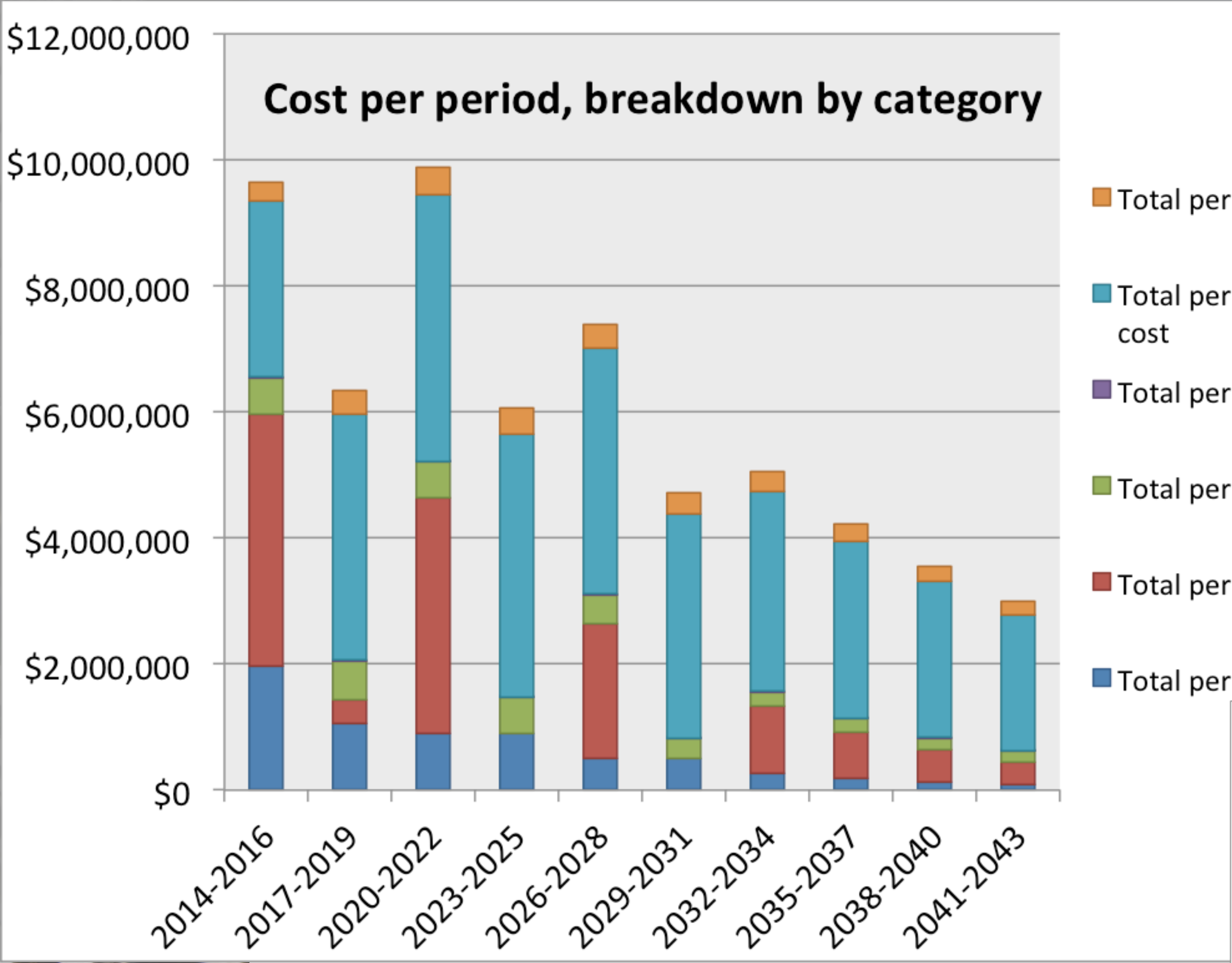
➢ **0.5 EB / year is probably an under estimate!**

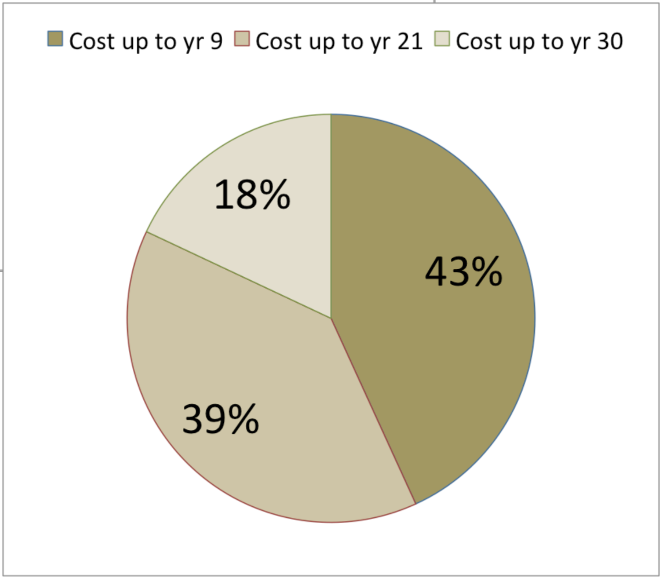
Cost Modelling: Regular Media Refresh + Growth

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



Case B) increasing archive growth





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

Balance sheet

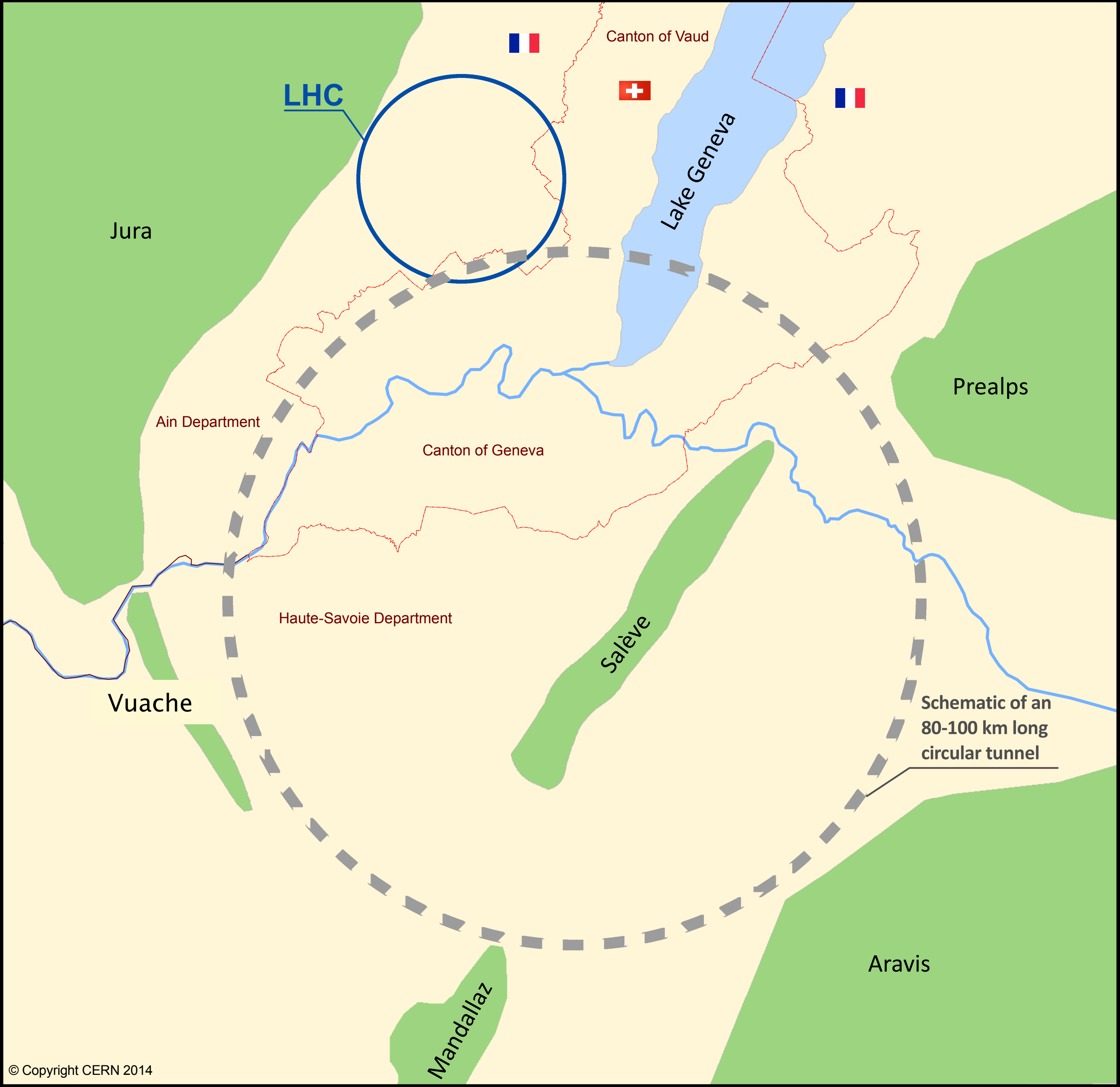
20 year investment in Tevatron			~ $4B
Students	$4B		
Magnets and MRI	$5-10B	}	~ $50B total
Computing	$40B		

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


LHC
Canton of Vaud
Lake Geneva
Jura
Prealps
Ain Department
Canton of Geneva
Haute-Savoie Department
Salève
Vuache
Schematic of an
80-100 km long
circular tunnel
Aravis
Mandallaz

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]

The Story So Far…

- We have invested heavily in many areas of sustainable, scalable, federated, reliable storage & services over many years

- We have a good understanding of the costs – a proven funding model + multi-decade outlook

- We should (could) feed this into on-going work to help build the 2020 / 2030 vision

Generic e-if vs VREs

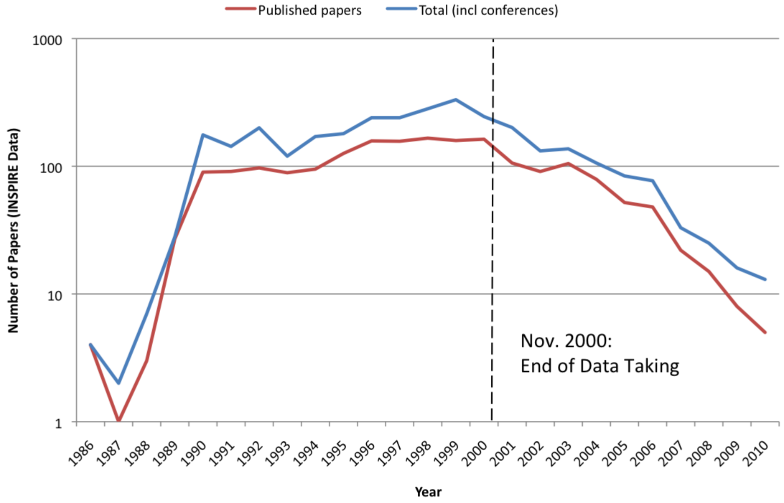
- We have also learned that attempts to offer "too much" functionality in the core infrastructure does not work (e.g. FPS)
- This is recognised (IMHO) in H2020 calls, via "infrastructure" vs "virtual research environments"
- There is a fuzzy boundary, and things developed within 1 VRE can be deployed to advantage for others (EGI-InSPIRE SA3)

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

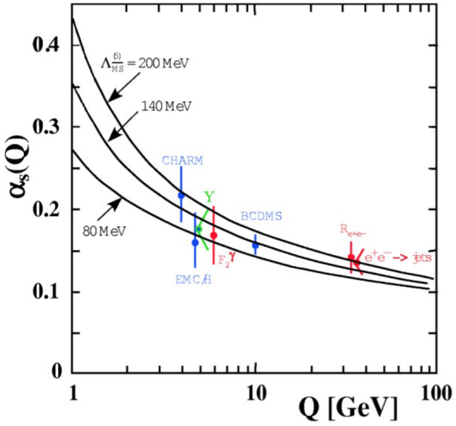
➢ **Agree with Funding Agencies clear targets & metrics**

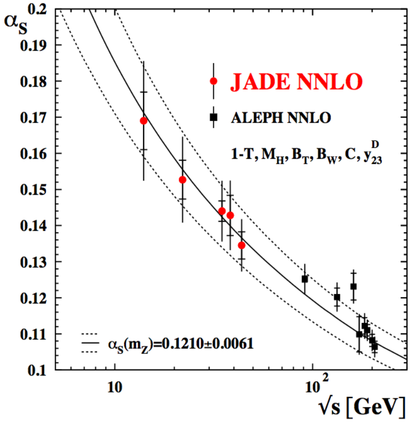
1 – Long Tail of Papers



Nov. 2000:
End of Data Taking

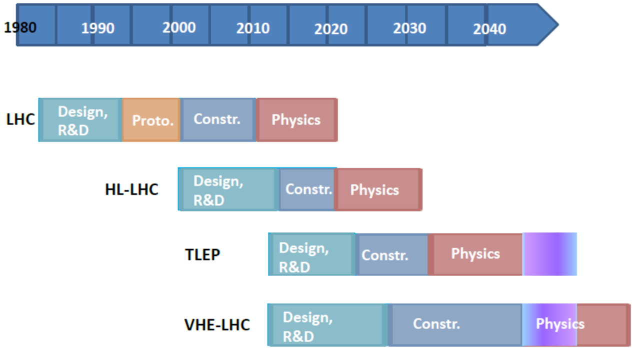
2 – New Theoretical Insights





3 – "Discovery" to "Precision"

possible long-term time line



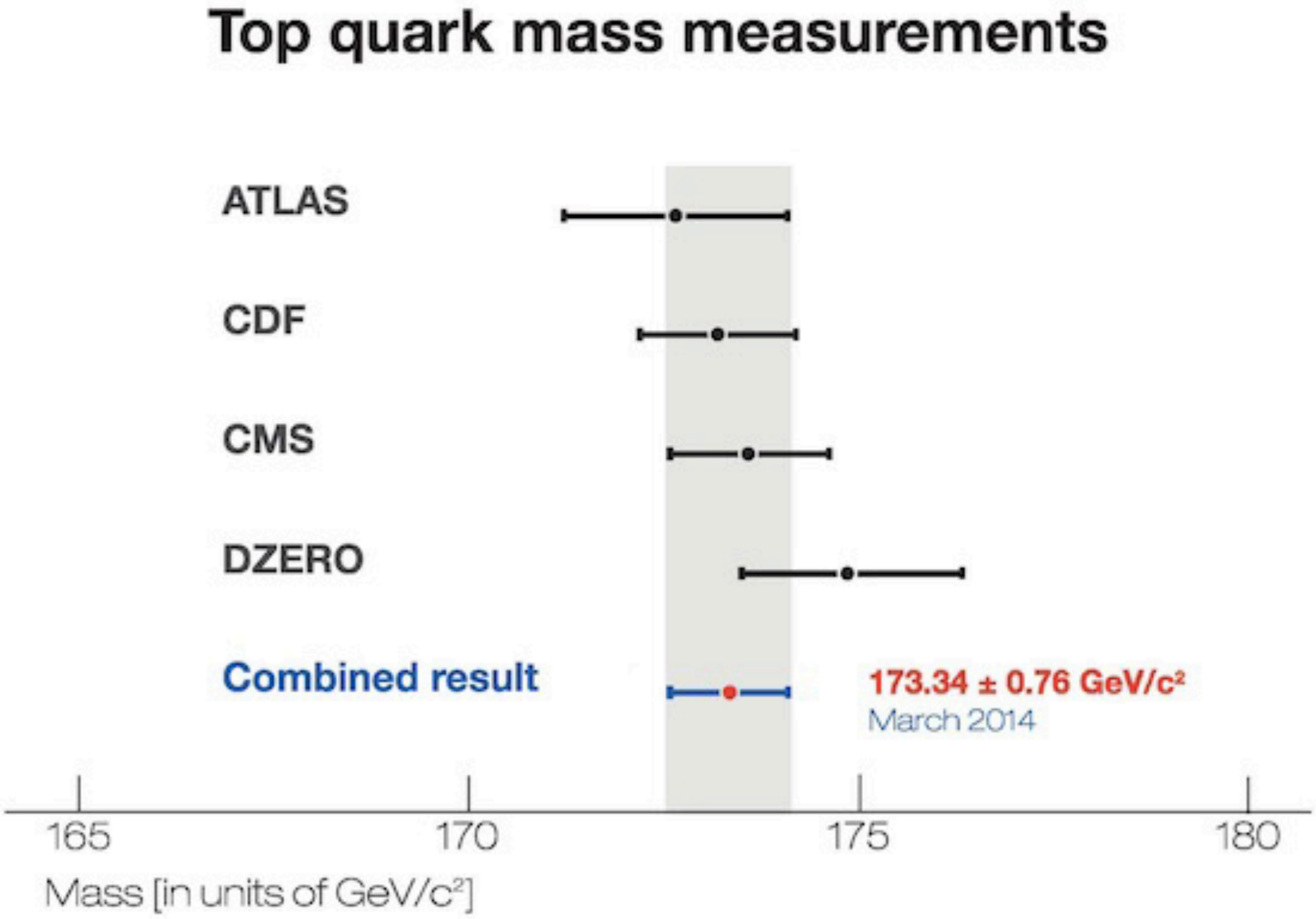
Zimmermann

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)

20 Years of the Top Quark



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)

How to respond?

a) **Each project / experiment responds to individual FA policies**
   - **n x m**

b) **We agree together – service providers, experiments, funding agencies – on a common approach**
   - **DPHEP can help coordinate**

- **b) almost certainly cheaper / more efficient but what does it mean in detail?**

Data Seal of Approval: 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.

Guidelines Related to Repositories (4-8):

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.
5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.
6. The data repository applies documented processes and procedures for managing data storage.
7. The data repository has a plan for long-term preservation of its digital assets.
8. Archiving takes place according to explicit work flows across the data life cycle.

Guidelines Related to Repositories (9-13):

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

10... enables the users to discover and use the data and refer to them in a persistent way.

11... ensures the integrity of the digital objects and the metadata.

12.… ensures the authenticity of the digital objects and the metadata.

13.The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Guidelines Related to Data Consumers (14-16):

14. The data consumer complies with access regulations set by the data repository.
15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.
16. The data consumer respects the applicable licences of the data repository regarding the use of the data.

DSA self-assessment & peer review

- Complete a self-assessment in the DSA online tool. The online tool takes you through the 16 guidelines and provides you with support
- Submit self-assessment for peer review. The peer reviewers will go over your answers and documentation
- Your self-assessment and review will not become public until the DSA is awarded.
- After the DSA is awarded by the Board, the DSA logo may be displayed on the repository's Web site with a link to the organization's assessment.

Collaboration - Benefits

- In terms of 2020 vision, collaboration with other projects has arguably advanced us (in terms of implementation of the vision) by several years

- **I typically quote 3-5 years and don't think that I am exaggerating**

- **Concrete examples include "Full Costs of Curation", as well as proposed "Data Seal of Approval+"**

- With or without project funding, we should continue – and even strengthen – this collaboration
  - APA events, iDCC, iPRES etc. + joint workshops around RDA

- **The HEP "gene pool" is closed and actually quite small – we tend to recycle the same ideas and "new ones" sometimes needed**

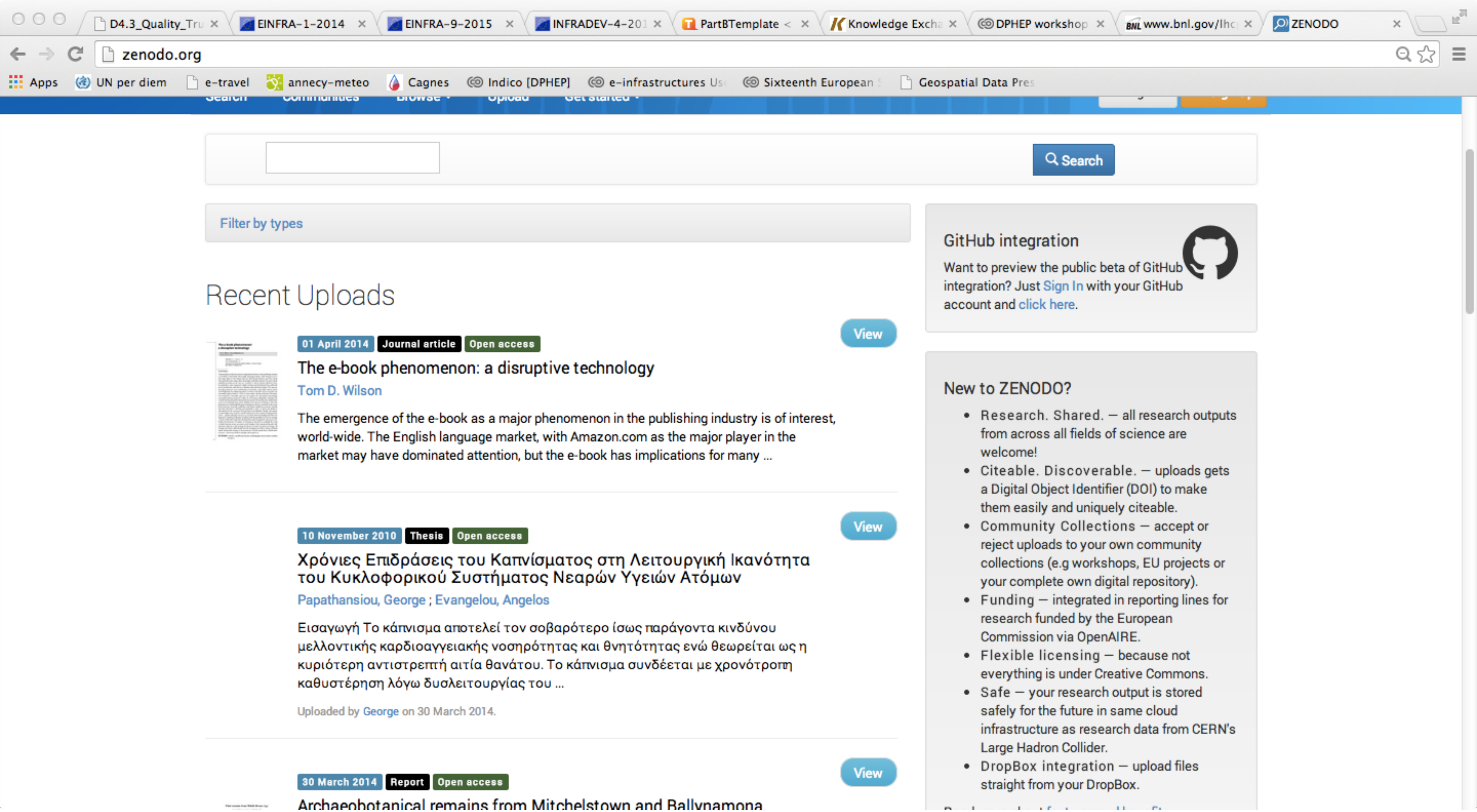
Additional Metrics (Beyond DSA)

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**

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



Recent Uploads

01 April 2014 | Journal article | Open access

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

Χρόνιες Επιδράσεις του Καπνίσματος στη Λειτουργική Ικανότητα του Κυκλοφορικού Συστήματος Νεαρών Υγειών 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

Archaeobotanical remains from Mitchelstown and Ballynamona

GitHub integration

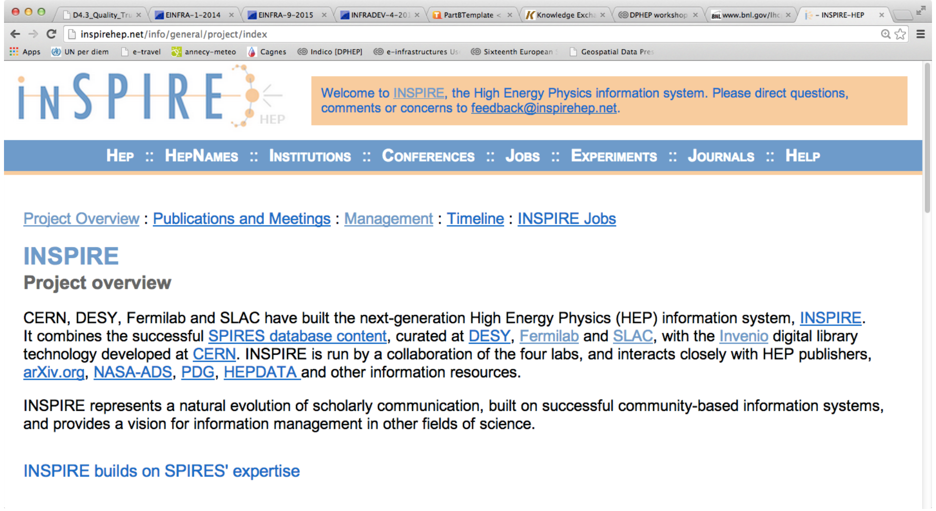
Want to preview the public beta of GitHub integration? Just Sign In with your GitHub account and click here.

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)



Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

Project Overview : Publications and Meetings : Management : Timeline : INSPIRE Jobs

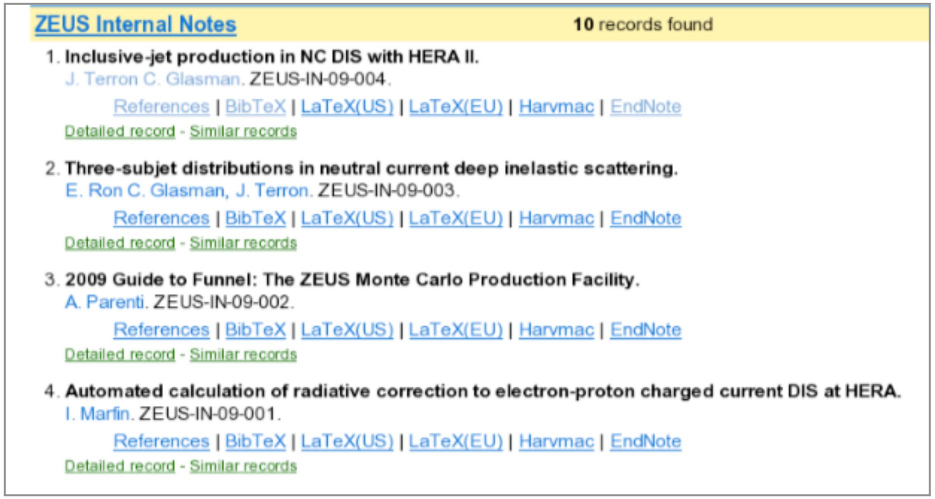
INSPIRE

Project overview

CERN, DESY, Fermilab and SLAC have built the next-generation High Energy Physics (HEP) information system, INSPIRE. It combines the successful SPIRES database content, curated at DESY, Fermilab and SLAC, with the Invenio digital library technology developed at CERN. INSPIRE is run by a collaboration of the four labs, and interacts closely with HEP publishers, arXiv.org, NASA-ADS, PDG, HEPDATA and other information resources.

INSPIRE represents a natural evolution of scholarly communication, built on successful community-based information systems, and provides a vision for information management in other fields of science.

INSPIRE builds on SPIRES' expertise



ZEUS Internal Notes **10** records found

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 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…**

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

Summary – Data Sharing

- **Together** we can do much more than if working alone

- Let’s **combine** our strengths and build the **future!**

BACKUP

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)

LHC Data Access Policies

Level (standard notation)	**Access Policy**
L0 (raw) (cf "Tier")	Restricted even internally • Requires significant resources (grid) to use
L1 (1^{st} processing)	Large fraction available after "embargo" (validation) period • Duration: a few years • Fraction: 30 / 50 / 100%
L2 (analysis level)	**Specific (meaningful) samples for educational outreach: pilot projects on-going** • **CMS, LHCb, ATLAS, ALICE**
L3 (publications)	**Open Access (CERN policy)**

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!