The Challenges of Long-Term Data Preservation

Data Preservation, Curation & Stewardship

Jamie.Shiers@cern.ch

Big Data Applications in Science & Industry

Budapest, March 2015

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

Introduction

- This presentation is based on the paper attached to the agenda
- It tells a story – how **collaboration** between experts in **different** disciplines can rapidly lead to a solution
    - Each “partner” comes with different technologies: working together we can solve problems faster, better and in a more sustainable fashion
    - The “**HEP gene pool**” is quite large (20K) but also quite closed…
- It finishes with some un-answered questions: areas for future work (& collaboration)

2020 Vision for LT DP in HEP

- ***Long-term – e.g. FCC 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**, (FAIRport?) through which data / tools accessed

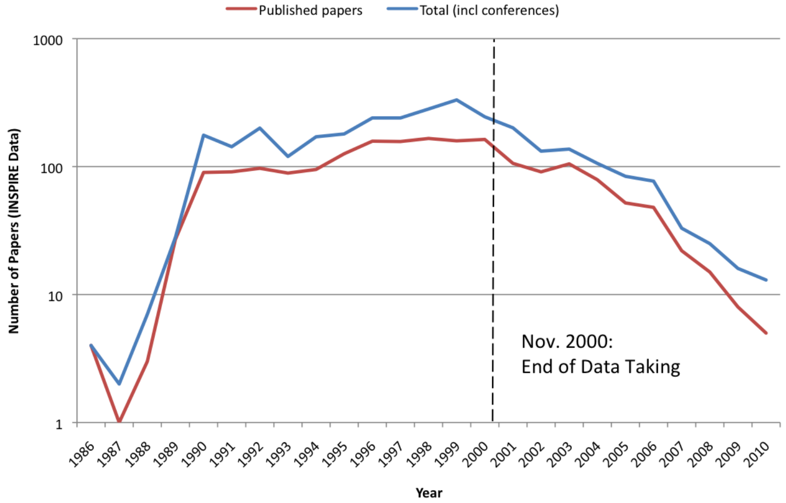
➢ **Agree with Funding Agencies clear targets & metrics**

No “One Size Fits All”

- DP: theory and practice driven by disciplines that make “observations”
  - By definition unrepeatable
- They have developed a set of ISO standards that are “adopted” by ~all disciplines WW
  - **There is a large amount of deep expertise, training materials and so forth – no “wheel to re-invent”**
- Other disciplines – e.g. public / private archives – have clearly defined “data destruction” policies

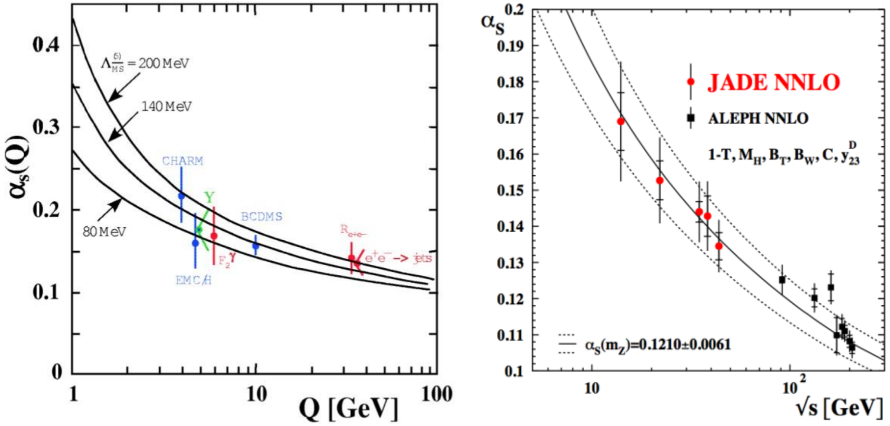
➢ **YOU** need to work out **WHAT** you want to keep and **WHY** – **WHO** will pay and **WHY**

1 – Long Tail of Papers



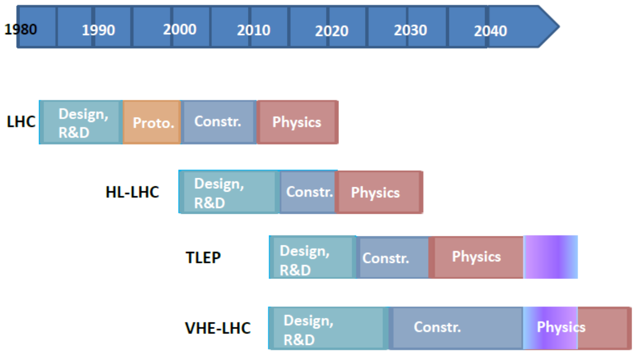
Nov. 2000:
End of Data Taking

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

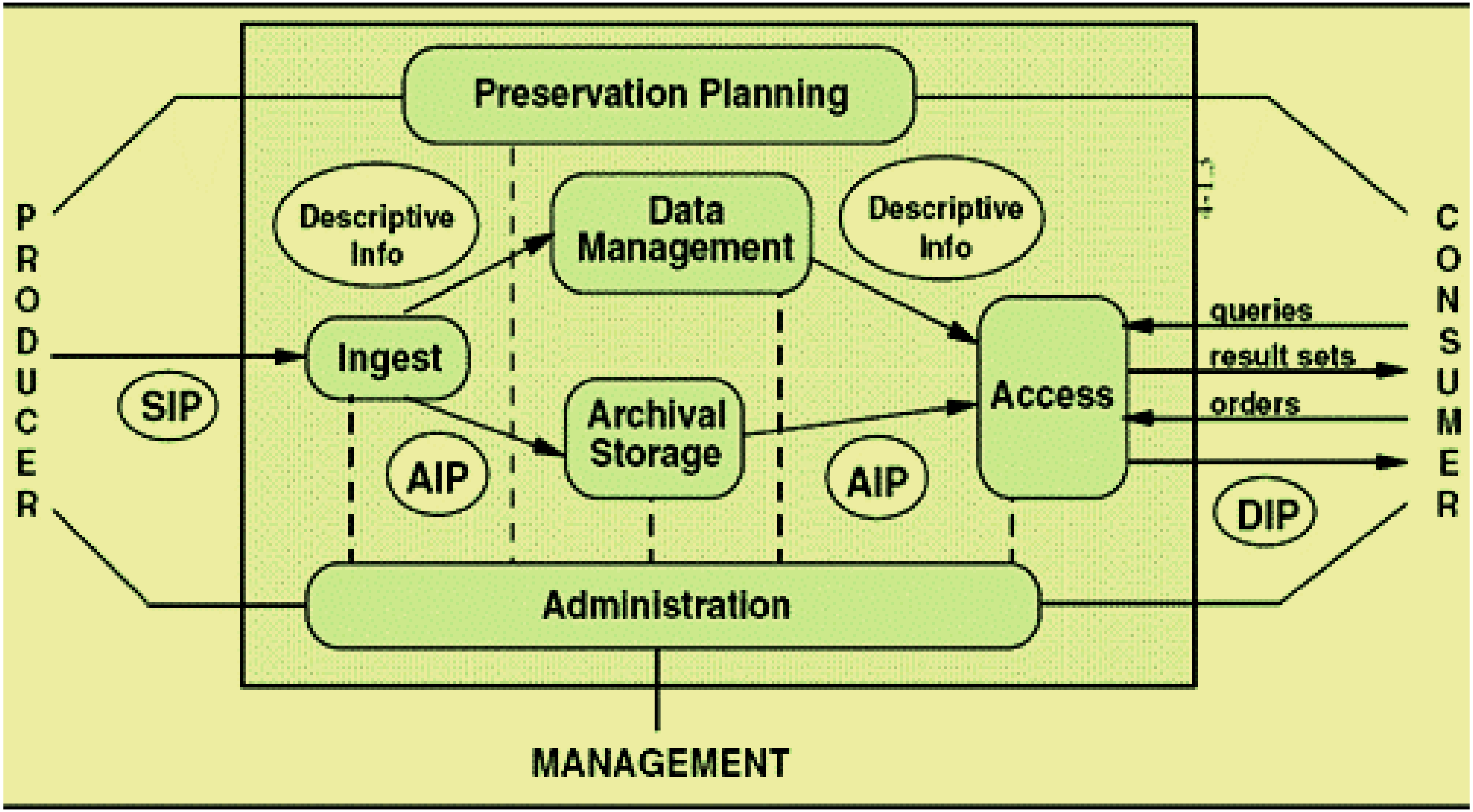
Volume: 100PB + ~50PB/year (+400PB/year from 2020)

4C Roadmap Messages

A Collaboration to Clarify the Costs of Curation

1. Identify the **value** of digital assets and make **choices**
2. Demand and choose more **efficient** systems
3. Develop **scalable** services and infrastructure
4. Design digital curation as a **sustainable** service
5. Make funding **dependent** on costing digital assets across the whole lifecycle
6. Be **collaborative** and **transparent** to drive down costs

OAIS – ISO 14721:2003



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.

DANS

Driven by data

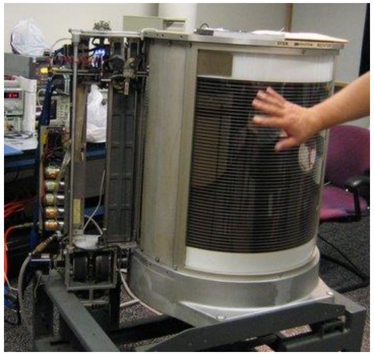
Guidelines Related to Repositories (4-8):

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 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.

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….

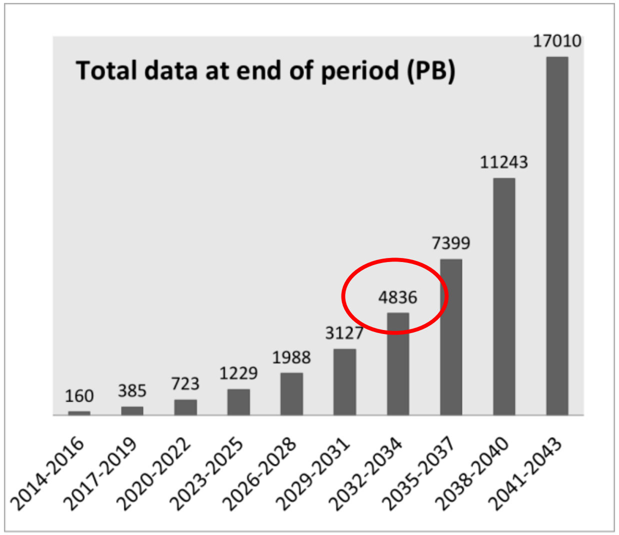
~300K tapes were 'archived'… .. ~150K were manually mounted….

..and then copied to Redwoods….

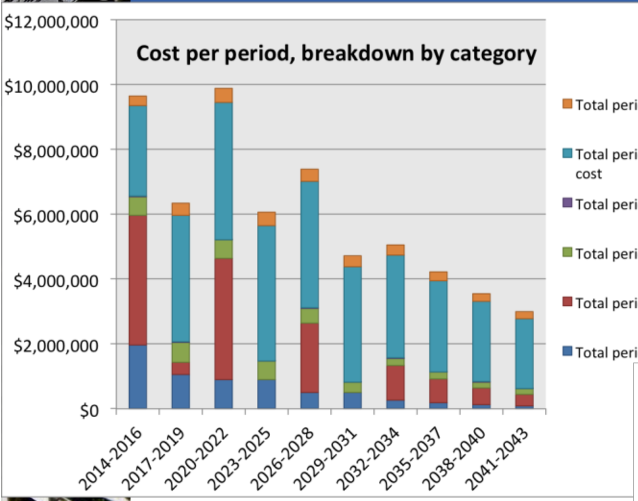
CERN

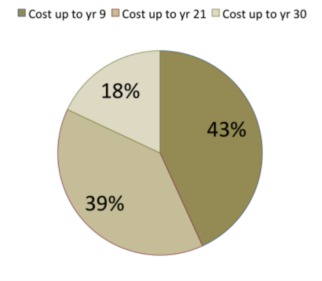
Cost Modelling: Regular Media Refresh + Growth

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



Case B) increasing archive growth





Total cost: ~59.9M$
(~2M$ / year)

Towards a CERN DP Strategy?

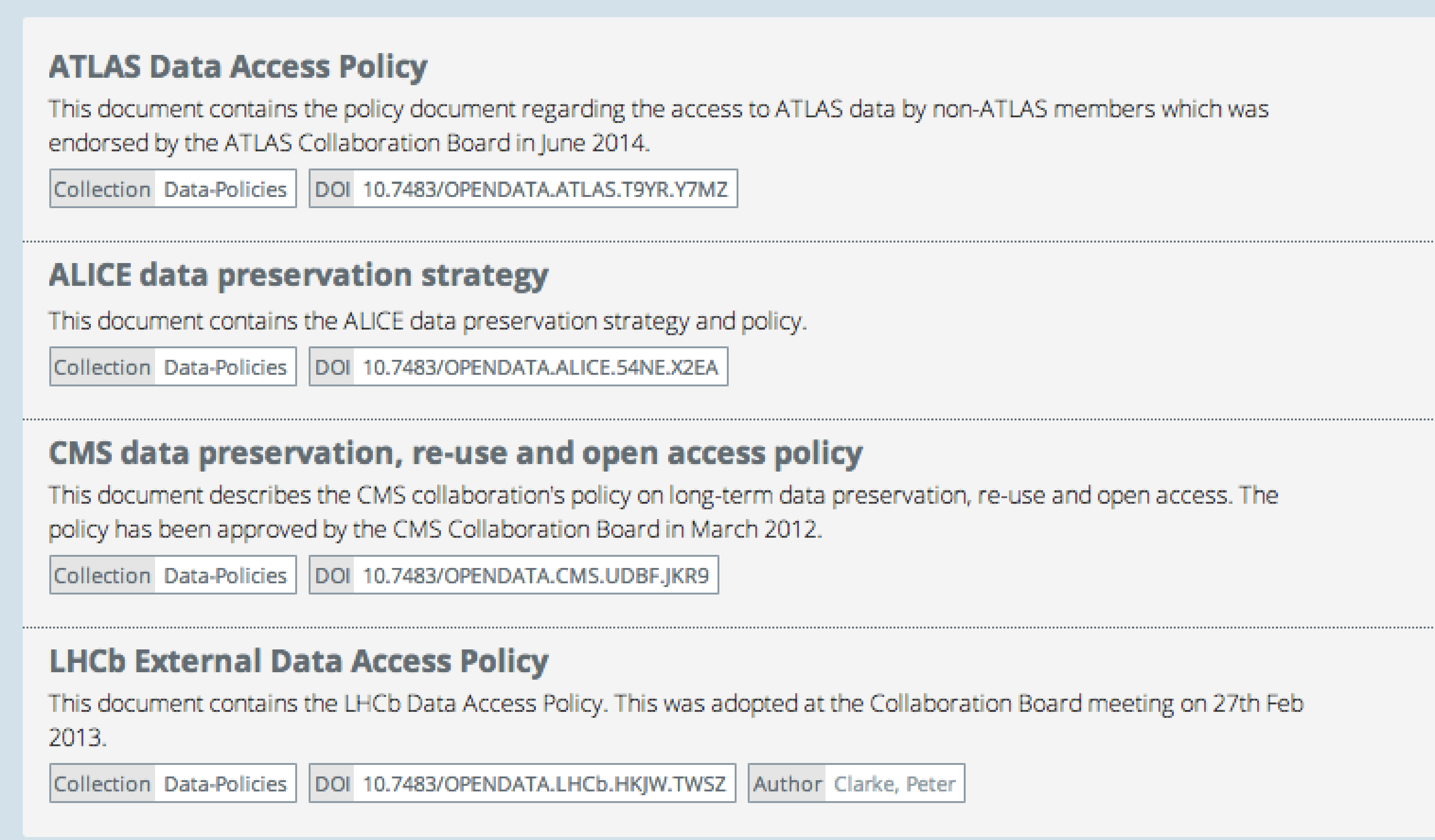
1. The updated Strategy for European Particle Physics, approved by Council in May 2014, states that *"**infrastructures for … data preservation … should be maintained and further developed**."*
2. Such infrastructures include ***digital repositories***, where ***copies*** or ***replicas*** of the data are kept.
3. As host laboratory, it is expected that (from now on?) a copy of all data acquired by CERN experiments *and* targeted for long-term preservation be stored in the CERN digital repository. […]
4. **It is strongly recommended that one or more copies of the above data are maintained outside, at or spread over institutes that form part of the collaboration.**

Use Cases – LHC (and LEP)

1. Preserve data, **software, and know-how** in the collaborations
2. Share data and associated **software** with larger scientific community – **O(PB) in 2020?**
3. Open access to reduced data sets to general public – **O(TB) ?**
4. Bit preservation (**O(100PB) today, 1EB ~2025, 10EB ~2035** – **ALREADY "FILTERED"**)

- Policies: http://opendata.cern.ch/collection/data-policies

http://opendata.cern.ch/collection/data-policies



ATLAS Data Access Policy

This document contains the policy document regarding the access to ATLAS data by non-ATLAS members which was endorsed by the ATLAS Collaboration Board in June 2014.

Collection Data-Policies DOI 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ

ALICE data preservation strategy

This document contains the ALICE data preservation strategy and policy.

Collection Data-Policies DOI 10.7483/OPENDATA.ALICE.54NE.X2EA

CMS data preservation, re-use and open access policy

This document describes the CMS collaboration's policy on long-term data preservation, re-use and open access. The policy has been approved by the CMS Collaboration Board in March 2012.

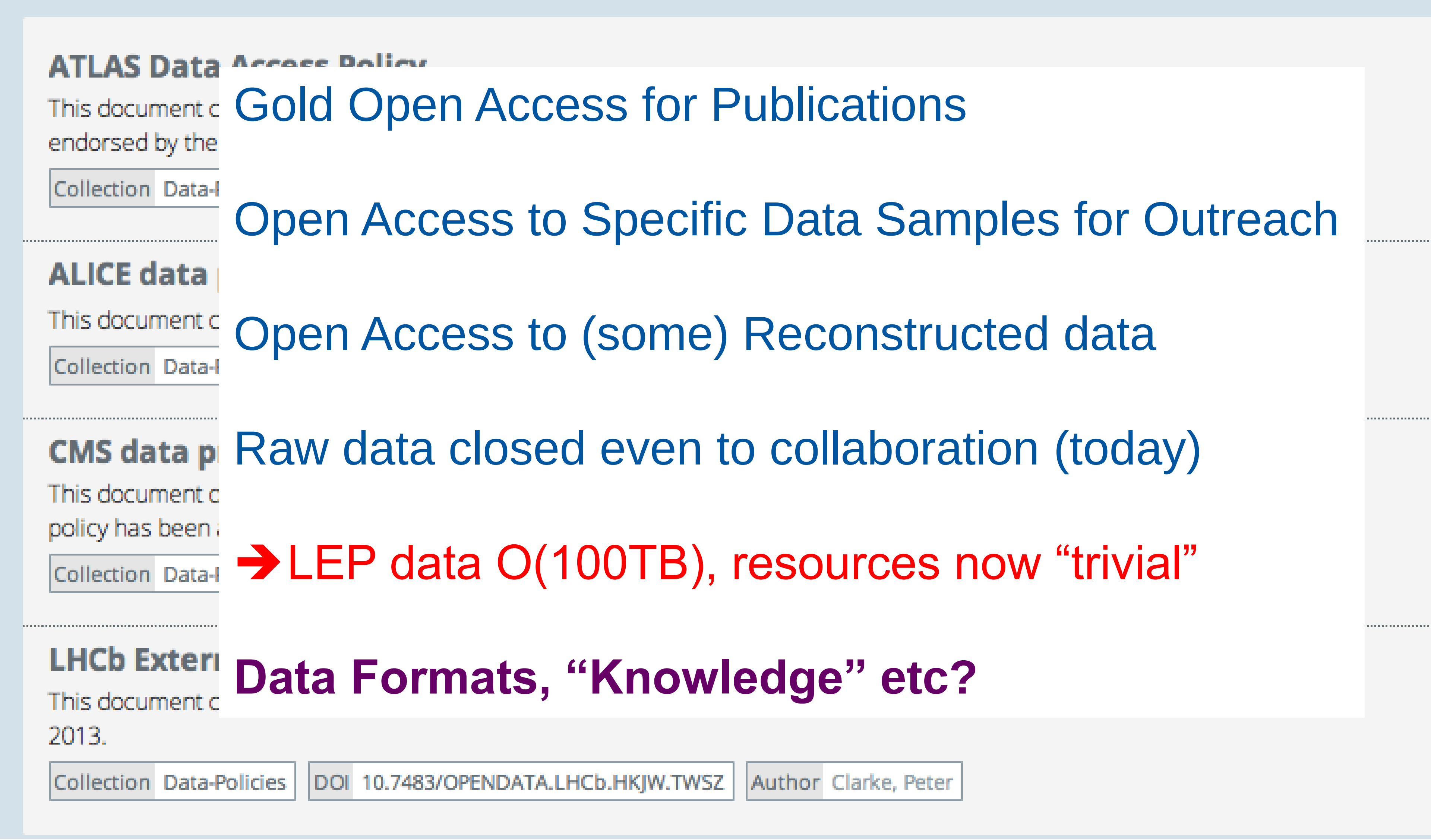
Collection Data-Policies DOI 10.7483/OPENDATA.CMS.UDBF.JKR9

LHCb External Data Access Policy

This document contains the LHCb Data Access Policy. This was adopted at the Collaboration Board meeting on 27th Feb 2013.

Collection Data-Policies DOI 10.7483/OPENDATA.LHCb.HKJW.TWSZ Author Clarke, Peter

http://opendata.cern.ch/collection/data-policies



ATLAS Data Access Policy

This document c

endorsed by the

Collection Data-

ALICE data

This document c

Collection Data-

CMS data p

This document c

policy has been

Collection Data-

LHCb Extern

This document c

2013.

Collection Data-Policies | DOI 10.7483/OPENDATA.LHCb.HKJW.TWSZ | Author Clarke, Peter

Gold Open Access for Publications

Open Access to Specific Data Samples for Outreach

Open Access to (some) Reconstructed data

Raw data closed even to collaboration (today)

➔LEP data O(100TB), resources now “trivial”

Data Formats, “Knowledge” etc?

2020 Vision for LT DP in HEP

- ***Long-term – e.g. FCC 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
    - **"HEP FAIRport": Findable, Accessible, Interoperable, Re-usable**
- **Agree with Funding Agencies clear targets & metrics**

http://science.energy.gov/funding-opportunities/digital-data-management/

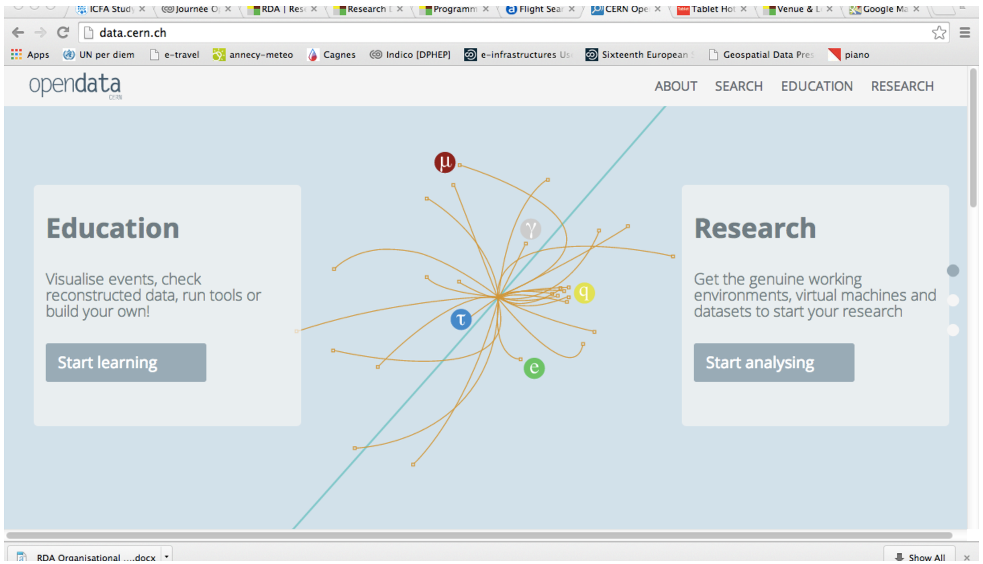
- ***"The focus of this statement is sharing and preservation of digital research data"***
- **All proposals submitted to the Office of Science (after 1 October 2014) for research funding must include a Data Management Plan (DMP) that addresses the following requirements:**

1. **DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved.**

If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4).

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.

U.S. DEPARTMENT OF ENERGY | Office of Science



Education

Visualise events, check reconstructed data, run tools or build your own!

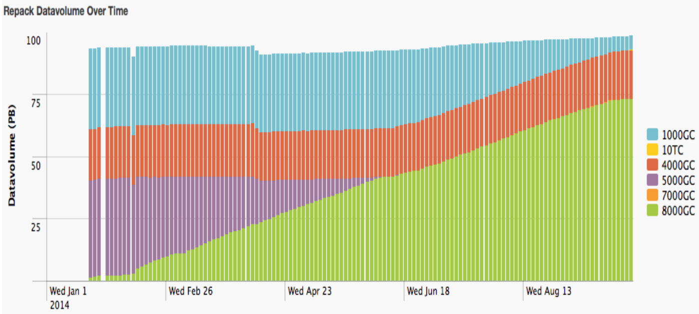
Research

Get the genuine working environments, virtual machines and datasets to start your research

DSS Repack

CERN IT Department

http://indico.cern.ch/event/CERN-ITTF-2014-09-26



- Oracle: Done
  - 39PB self-repacked (5->8TB), 27PB 1TB emptied
- IBM: Dec'14-Mar'15
  - 20PB of IBM 4TB to self-repack and 5.6PB 1TB tapes to empty
- All repacked media has been verified
- All problem source tapes identified and being handled (cf next slides)
- Cleanup of tape pools and (properly) establishing double copies
  - across buildings
  - complete second copies where missing (ie OPAL)

DP – ‘The Big Rocks”

1. **The Data itself: don’t try to do it alone – don’t try to do it at home: Scale & Sustainability**

2. **The Business Case – this will (probably) be specific to your domain; your Use Cases; but sharing with others will help**

3. **“Knowledge capture & preservation”**

- ***Not* specific tools, such as portals, digital libs etc.**

The Challenge(s)

1. **Reproducibility of results – over long periods of time and changing e-infrastructures**

2. **Data Sharing – even with long-ish embargo periods – can translate to significant demands**

3. **From Open Access to Open Data to Open Knowledge**

Some Questions re Open Data

- **The volumes involved – at least for HEP – could reach many PB or even beyond.**

- Who will pay? (Cost Recovery Patterns in Research Data Repositories)

  - ***Is it financially affordable?***
  - ***Is it technically implementable?***
  - ***Is it scientifically (or educationally, or culturally) meaningful?***

- The answers to these questions may well vary with time (see LEP) and also depend on the implementation(s) that we choose:

- **Open Access is just one step in the progression towards Open Data and finally “Open Knowledge”.**

2020 Vision for LT DP in HEP

- *Long%erm*–*e.g.*FCC*/ mescales:"disrup/ ve*change**
  - By 2020, all **archived)data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated) communi4es** 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
    - **"HEP)FAIRport":)Findable,)Accessible,)Interoperable,)Re‐usable)**

Agree)with)Funding)Agencies)clear)targets)&)metrics)

http://science.energy.gov/funding-opportunities/digital-data-management/

- *"The focus of this statement is sharing and preservation of digital research data"*
- **All proposals submitted to the Office of Science (after 1 October 2014) for research funding must include a Data Management Plan (DMP) that addresses the following requirements:**

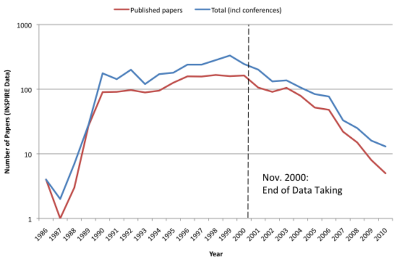
1. **DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved.**

   If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4).

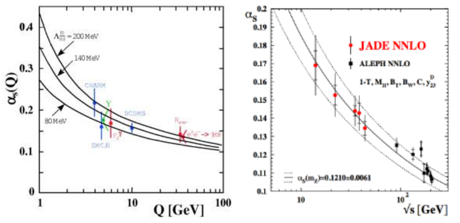
   At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

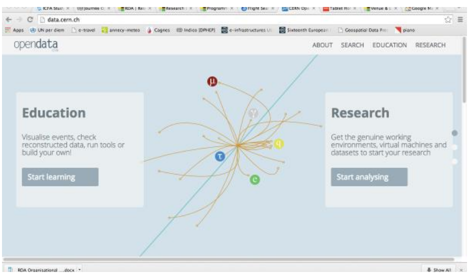
U.S. DEPARTMENT OF ENERGY | Office of Science

1"–"Long"Tail"of"Papers"



2"–"New"Theore+cal"Insights"





Education

Visualise events, check reconstructed data, run tools or build your own!

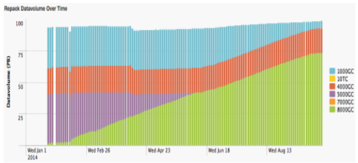
Research

Get the genuine working environments, virtual machines and datasets to start your research

DSS Repack

CERN IT Department

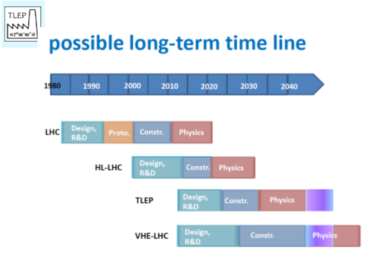
http://indico.cern.ch/event/CERN-ITTF-2014-09-26



- Oracle: Done
  - 39PB self-repacked (5->8TB), 27PB 1TB emptied
- IBM: Dec'14-Mar'15
  - 20PB of IBM 4TB to self-repack and 5.6PB 1TB tapes to empty
- All repacked media has been verified
- All problem source tapes identified and being handled (cf next slides)
- Cleanup of tape pools and (properly) establishing double copies
  - across buildings
  - complete second copies where missing (ie OPAL)

3"–"“Discovery”"to"“Precision”"

possible long-term time line



Zimmermann

Alain Blondel TLEP design study r-ECFA 2013-07-20

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
(+500PB/year from 2025)

4C Roadmap Messages

A Collaboration to Clarify the Costs of Curation

1. Identify the **value** of digital assets and make **choices**
2. Demand and choose more **efficient** systems
3. Develop **scalable** services and infrastructure
4. Design digital curation as a **sustainable** service
5. Make funding **dependent** on costing digital assets across the whole lifecycle
6. Be **collaborative** and **transparent** to drive down costs

Balance'sheet'–'Tevatron@FNAL'

- 20 year investment in Tevatron ~'$4B'
- Students $4B
- Magnets and MRI $5-10B } ~'$50B'total'
- **Compu>ng' ' '$40B''''**

Very%ough%calcula- on%%but%onfirms%our%gut%feeling%hat% investment%n%fundamental%cience%pays%ff%%

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

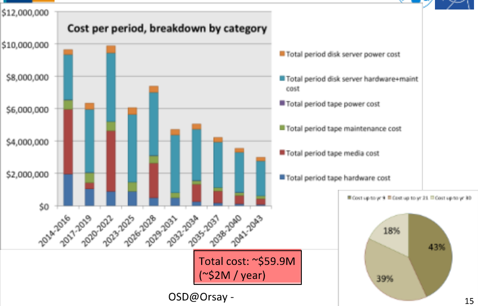
cf. STFC study of SRS Impact

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

Science & Technology Facilities Council

Approximation of (HL-)LHC Growth

HEPiX DESY CERN



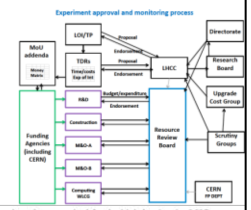
Total cost: ~$59.9M (~$2M / year)

Sustainability+–+Funding++

ORGANISATION EUROPEENNE POUR LA RECHERCH
CERN EUROPEAN ORGANIZATION FOR NUCLE

Action to be taken

For Approval	EUROPEAN STRATEGY SESSION OF COUNCIL 16th Session - 30 May 2013 European Commission Berlaymont Building - Brussels



i) The success of particle physics experiments, such as those required for the high-luminosity LHC, relies on innovative instrumentation, state-of-the-art infrastructures and large-scale data-intensive computing. *Detector R&D programmes should be supported strongly at CERN, national institutes, laboratories and universities. Infrastructure and engineering capabilities for the R&D programme and construction of large detectors, as well as infrastructures for data analysis, data preservation and distributed data-intensive computing should be maintained and further developed.*

What Next?

- **Training on, and certification of, sites as "Trusted Digital Repositories"**
- **Expanding "DPHEP Portal" to other (non-LHC) experiments and external sites**
- **Supporting key experiment Use Cases / Funding Agency Requirements**
  - **Reproducibility, Open Access for Outreach, DMPs**
- ➢ **Ensuring everything is sustainable, documented, "standards-based" and complete**