☺ **Data Preservation Update** ☺
DPHEP w/s, RDA, H2020, APA, …

Jamie.Shiers@cern.ch

February 2014

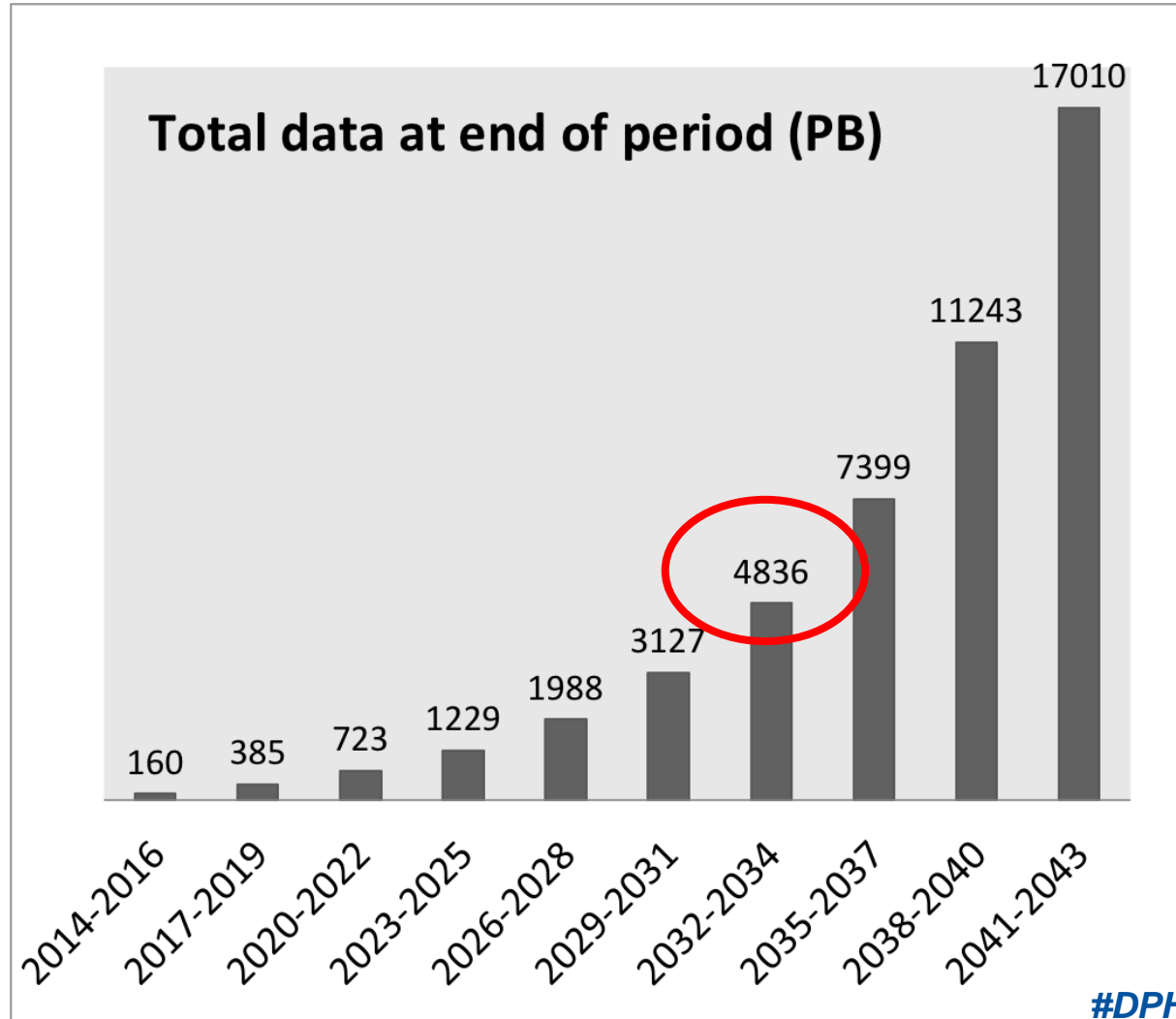International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics
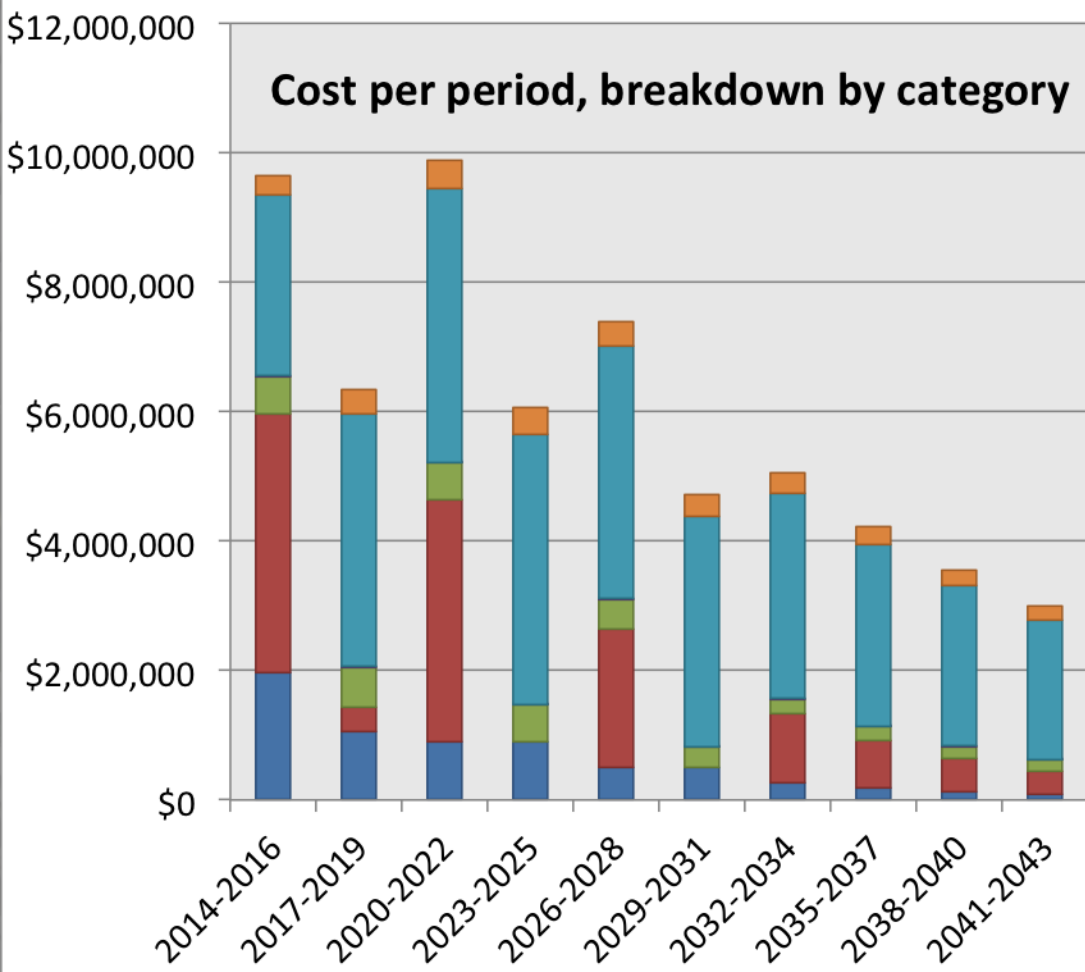
Executive Summary

- Things are going well.

- And will soon be going better.

DPHEP FCC: Workshop Goals

- The **primary goal** of this workshop [was] to understand the "Full Costs of Curation" ("in Context") for HEP data over 1 / 2 / 3 decades
  - @CERN: 100PB growing to a few (5?) EB, dominated by LHC

- Short term goal: data preservation costs **as part** of Medium Term Plans (MTP) **[Sergio Bertolucci: "to ensure continuity"]**

- Also include in **RRB** process: budgeted (or "need"); reviewed

? **How can this be used by other sites / experiments ?**

CERN **IT** Department

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

Total data at end of period (PB)

Period	PB
2014-2016	160
2017-2019	385
2020-2022	723
2023-2025	1229
2026-2028	1988
2029-2031	3127
2032-2034	4836
2035-2037	7399
2038-2040	11243
2041-2043	17010

#DPHEP

CERN**IT** Department
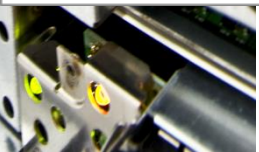
Cost per period, breakdown by category

Legend:
- 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

Pie chart legend: Cost up to yr 9 / Cost up to yr 21 / Cost up to yr 30

43% / 39% / 18%
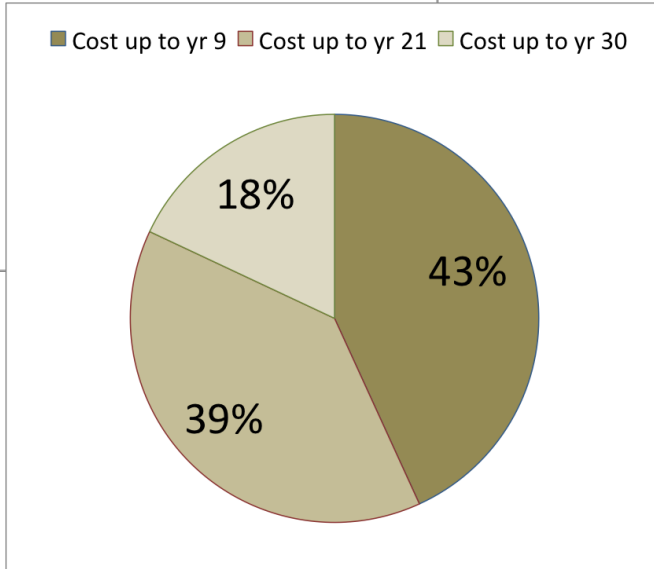
Total cost: ~$59.9M (~$2M / year)

#DPHEP

Costs and Scale (DPHEP@APA)

- There are 4 (main) collaborations + detectors at the LHC: the largest has **3000 members[1]**

- The annual cost of WLCG (infrastructure, operations, services) is **~EUR100M**

- The CERN database services cost around **2MCHF** per year for Materials (licenses, maintenance, hardware) – and **2MCHF** for personnel

- The central grid Experiment Integration Support team varied between **4-10** people, plus significant effort at sites and within experiments

- The DPHEP Full Costs of Curation workshop concluded that a team of **~4** people, with access to experts, could "**make significant progress**" (be careful with this number – it is **not enough** for the work-plan (next)!)

    [1] – source: http://atlas.ch/what_is_atlas.html#2a

1. DPHEP Portal

2. **Digital library** tools (Invenio) & services (CDS, INSPIRE, ZENODO) + related tools (HepData, RIVET, ...)

3. **Sustainable software**, coupled with advanced **virtualization** techniques, "snap-shotting" (incl. CernVM[FS]) ***and*** **validation** frameworks

4. Proven bit preservation at the 100PB scale, together with a **sustainable** funding model with an outlook to 2040/50 (+HEPiX coordination + RDA WG + H2020?)

5. *Open Data ("Open everything")*

In Terms of Projects…

1. **DPHEP portal: build in collaboration with other disciplines, including RDA IG and the APA…**
   – **Proposal being prepared**
2. **Digital libraries: continue existing collaborations & projects (also H2020 funding…)**
   – **Services in place and in use**
3. **Sustainable "bit preservation" – certified repositories as part of EINFRA-1-2014 – based on RDA WG "RFCs" for interfaces, functionality, federation etc.**
   – **Also a service! Certification and standardisation desirable…**
4. **"Knowledge capture & preservation": collaborate with SCIDIP-ES, APA(RSEN) and many others**
5. **Open "Big Data": a Big Opportunity (for RDA?)**
   – **Proposal exists – to be discussed in DPHEP shortly & with EU JRC / EUDAT / RECODE**

E-Infrastructures

EINFRA-1-2014 – Managing, preserving and computing with big research data (Sept'14)

(1) Establishing a federated pan-European data e-infrastructure to provide cost-effective and interoperable solutions for data management and long term preservation. EUDAT (including exascale bit preservation)

(2) Services to ensure the quality and reliability of the e-infrastructure, including certification mechanisms for repositories and certification services to test and benchmark capabilities in terms of resilience and service continuity of e-infrastructures. ISO 16363 standardisation related to bit data preservation. Standard based e-infrastructure monitoring as part of Helix Nebula marketplace.

(3) Federating institutional and, if possible, private data management and curation tools and services used across or at some point of the full data lifecycle, including approaches for identification of open data sources and data collected with sensitive or restricted access features. Helix-Nebula + ReAcH?

(4) Large scale virtualisation of data/compute centre resources to achieve on-demand compute capacities, improve flexibility for data analysis and avoid unnecessary costly large data transfers. Helix-Nebula + ReAcH

(5) Development and adoption of a standards-based computing platform (with open software stack) that can be deployed on different hardware and e-infrastructures (such as clouds providing infrastructure-as-a-service (IaaS), HPC, grid infrastructures…) to abstract application development and execution from available (possibly remote) computing systems. This platform should be capable of federating multiple commercial and/or public cloud resources or services and deliver Platform-as-a-Service (PaaS) adapted to the scientific community with a short learning curve. Adequate coordination and interoperability with existing e-infrastructures (including GÉANT, EGI, PRACE and others) is recommended Helix-Nebula marketplace working with EUDAT (1), EGI (6), PRACE, GEANT, OPENAIRE

(6) Support to the evolution of EGI (European Grid Infrastructure) towards a flexible compute/data infrastructure capable of federating and enabling the sharing of resources of any kind (public or private, grid or cloud, etc.) in order to offer computing and storage services to the whole European scientific community. The proposal will address operations for supplying services (IaaS, PaaS, SaaS) at European level, engagement of and tailoring of services to new user communities and dissemination activities. EGI transition to Helix-Nebula + ReAcH. Max 8M€

(7) Proof of concept and prototypes of data infrastructure-enabling software (e.g. for databases and data mining) for extremely large or highly heterogeneous data sets scaling to zetabytes and trillion of objects. Clean slate approaches to data management targeting 2020+ 'data factory' requirements of research communities and large scale facilities (e.g. ESFRI projects) are encouraged. Consider EOS + Ceph with Hi-LHC (and SKA + other ESFRI?) as use cases. Higher-level services, workflows. Work with ATLAS, CMS, Russia & NSF

(8) Enable the creation of a platform and infrastructure for mining text aggregated from different sources/publishers that responds to the needs of users (researchers). Zenodo + CERN library could potentially participate as a partner???

Giovanni Lamanna EU-T0 paper could be HEP-APPEC community angle

E-Infrastructures

EINFRA-1-2014 – Managing, preserving and computing with big research data (Sept'14)

(1) Establishing a federated pan-European data e-infrastructure to provide cost-effective and interoperable solutions for data management and long term preservation. EUDAT (including exascale bit preservation)

(2) Services to ensure the quality and reliability of the e-infrastructure, including certification mechanisms for repositories and certification services to test and benchmark capabilities in terms of resilience and service continuity of e-infrastructures. ISO 16363 standardisation related to bit data preservation. Standard based e-infrastructure monitoring as part of Helix Nebula marketplace.

(3) Federating institutional and, if possible, private data management and curation tools and services used across or at some point of the full data lifecycle, including approaches for identification of open data sources and data collected with sensitive or restricted access features. Helix-Nebula + ReAcH?

Actions

- Launch + drive RDA WG on bit preservation at March RDA P3
- H2020 project including "bit preservation" services -> sustainability: now-Sep 2014 (funds from 2015 if ok)
- Strengthen multi-directional collaboration with APA(RSEN), SCIDIP-ES, EUDAT in H2020 and outside
  - Plus many other projects worldwide
- Agree on functionality of DPHEP portal; prototype
- CMS-CSC pilot using CernVM[FS], including metrics
- Open Data – pursue opportunity through RDA, ERF workshop in Brussels: timescale 2015+
- (Self-) audit? (OAIS + ISO 16363 + RDA + …)
- *Establish "core team" of ~4 people in 2014?*

S.W.O.T.

Strengths	Clear break-down of work-plan, good understanding of costs & benefits, commitment from service providers: **PH-SFT, IT-CIS, IT-DI, IT-DSS, GS-SIS, …**
Weaknesses	(Some) needed manpower missing: chicken & egg? (Successful project proposals should provide necessary impetus; IMHO strong message from experiments + management bodies would also help) **[See summary]** ☺ **ICFA statement + ESPP + CERN Council Chair + DG**
Opportunities	Clear opportunities for peer-peer collaboration – and funding – with numerous projects and communities **(!!! We still have a lot to learn !!!)**
Threats	Do we have commitment from funding agencies – backed by experiments – of this plan (Some FAs: **YES!**) **Do non-CERN experiments / sites agree?**

External Collaboration

- People / projects are extremely interested in what we are doing
- They are keen to collaborate – and contribute
- There are numerous events – sometimes clashing – in different parts of the world
  - See DPHEP Indico pages
- We should ensure a presence at least at the key events – including also more technical aspects of the work, and results

Summary

- In several key areas, there are **services** and **solutions** in place now that can be used
- In others, there are **opportunities** to address these in the short or medium terms via open collaborations – funding also possible
- **Long-term sustainability** foreseen in plans – should ensure that preservation issues are included in formal planning (MTP, RRB, etc.)
- This will also involve **formal approval** of funding agencies, experiments and "management"

Executive Summary Revisited

- Things are going very well.

- And will soon be going better.

➢ **Discussion of needed resources at e.g. RRB will provide necessary clarification!**

✓ **This should also include break-down of existing effort and allow that to be formally recognised!**