Topical Workshop: Costs of Curation

Preparing a Business Plan for DPHEP

Jamie.Shiers@cern.ch

January 2014

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

GOALS

Workshop Goals

- The **primary goal** of this workshop is to understand the "Full Costs of Curation" for HEP data over 1 / 2 / 3 decades
  - @CERN: 100PB growing to a few EB, dominated by LHC

- Short term goal: data preservation costs **as part** of Medium Term Plans (MTP)
  - CERN GS, IT, PH: ***wanting, willing and waiting to help…***
- Also include in **RRB** process: budgeted (or "need"); reviewed

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

? To Answer This !

- A central message of the DPHEP blueprint is that *data preservation in HEP is not possible without* ***long term investment*** *in not only hardware (M) but also* ***human resources (P)***

- **For (any) funding, we must explain exactly how much M + P we need, how this evolves with time, what the benefits are etc.**

Setting the Scale

- We know the annual cost of WLCG
  - Around EUR100M integrated over T0/T1/T2
- We know the CERN budget, the cost of the LHC

- ***A "few per cent" of the WLCG costs would not appear unreasonable to me..***

- How does this evolve with time?

#DPHEP

SERVICES

HEP Preservation in Practice

IMHO, we should now be talking about:

- **Sustainable services, solutions and support:**
  - **Coordinated** via the DPHEP Implementation Board
  - Which includes inter-laboratory / cross-experiment **collaboration & projects**
  - Collaboration with **other disciplines** through existing fora, e.g. **APA** in its projection into **RDA** space

The Approach (DPHEP@RDA-2)

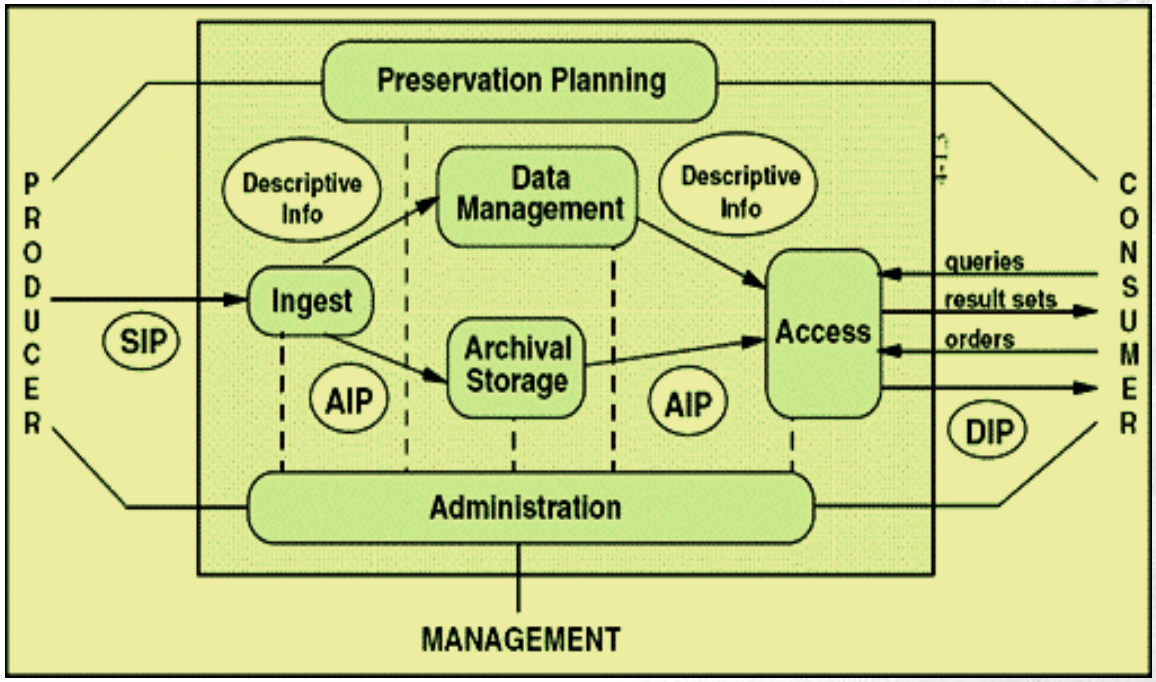
- Whilst retaining a holistic view, the problem is broken down into a number of key areas. Each is addressed using state-of-the-art techniques, that include:

1. **Digital library** tools (Invenio[2]) & services (CDS[3], INSPIRE[4], ZENODO[5]) + related tools (HepData, RIVET, …)
2. **Sustainable software**, coupled with advanced **virtualization** techniques[6] and **validation** frameworks[7]
3. Proven bit preservation at the 100PB scale, together with a **sustainable** funding model with an outlook to 2040/50

BRIDGES

Digital Preservation Solution

- the ISO OAIS Reference Model for *an* OAIS. This reference model is defined by recommendation CCSDS 650.0-B-1 of the Consultative Committee for Space Data Systems;[1] this text is identical to ISO 14721:2003.



Source: Long-Term Preservation of Digital Documents. 2006. doi:10.1007/978-3-540-33640-2. ISBN 978-3-540-33639-6.
Public Domain.

Digital processes break easily

- Short-period funding
- Software lifecycle: code, interfaces, formats…
- Dependent on expert knowledge
- Thin documentation and metadata

Source: : wikipedia PD Image resources

Bridging Components

- We need to design and build “bridges” between the individual components
- Respecting an overall architecture (OAIS)
- Ensuring the implementation is not tied to “CERN services” – e.g. fully applicable to other HEP experiments / services at other sites
- **As “future-proof” as possible**
- We need to do this **together** – and profiting from extensive existing experience “elsewhere”

➢ **A “hurried implementation” now could cost a huge amount in the long term**

Summary

- After several years of study and analysis, the DPHEP Study Group delivered a **Blueprint**
- A summary was input to the ESPP update and **Data Preservation** is now **part** of the European Strategy for Particle Physics
- A **small** set of services / projects have been identified / agreed upon
- These, together with the **associated resources**, should now be considered – and handled – as **FULL PRODUCTION SERVICES**

RECAP

Workshop Goals

- The primary goal of this workshop is to understand the "Full Costs of Curation" for HEP data over 1 / 2 / 3 decades
  - 100PB growing to a few EB
- We have to base our estimates on "reasonable" assumptions and scenarios

- Short term goal: data preservation costs as part of Medium Term Plans (MTP)
  - CERN GS, IT, PH: ***wanting, willing and waiting to help…***
- Also include in RRB process

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

#DPHEP