DP (C)HEP

Sustainable Strategies for Long-Term DP at the Exa-scale

Jamie.Shiers@cern.ch

CHEP 2013 - Amsterdam

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

Overview

- This presentation will give an update on DPHEP since ~CHEP 2012, together with some directions
- This includes progress on (some of) the recommendations in the DPHEP Blueprint
- There are numerous DP-related talks / posters during CHEP
- Plus a dedicated session on Wednesday pm
- See also recent WLCG GDB presentation on Use & Business Cases, Costs and Cost Models

DPHEP Progress / Status

- Blueprint summary was submitted to the ESPP update
- Data preservation is now officially part of the revised strategy (which obviously focusses on physics)
- A DPHEP Project Manager is now in place: 3 year term (2013 – 2015)
- **Moving from a Study Group to a Collaboration**

➢ **Now is the time to define concrete projects / services: those in Blueprint and beyond**

- **Funding: host labs, institutes, national, EU, more**

DPHEP Progress / Status

- Blueprint summary was submitted to the ESPP update
- Data preservation is now officially part of the revised strategy (which obviously focusses on physics)
- A DPHEP Project Manager is now in place: 3 year term (2013 – 2015)
- **Moving from a Study Group to a Collaboration**

➢ **Now is the time to define concrete projects / services: those in Blueprint and beyond**

- **AND INCLUDE COSTS!**

- **Funding: host labs, institutes, national, EU, more**

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 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

➢ **Vision achievable, but we are far from this today**

Data Preservation Maturity Model

Level	Metric	Implications
4	Reproducible results by "citizen scientists"	Desired(?) by funding agencies: people able to reproduce an analysis should be awarded "a degree" – beyond what can realistically be afforded?
3	Reproducible results where consumer ≠ producer and outside immediate community	Stronger demonstration of long-term preservation. Knowledge stored is sufficient for physicist outside immediate community to reproduce results
2	**Reproducible results where consumer ≠ producer but within same "larger community", e.g. LHC (ATLAS / CMS; CDF / D0, …)**	**Highly desirable for "minimal" long-term preservation. "Knowledge" stored is sufficient for a physicist from a different collaboration (but within same overall programme) to reproduce results**
1	Reproducible results where consumer = producer	Required during lifetime of collaboration
0	N/A	Data is lost: logically or physically. This is probably the reality for the bulk of pre-DPHEP experiments (and even some of those??)

- Scale (complexity) is probably "exponential"

Software Preservation Maturity Model

Level	Metric	Implications
4	Reproducible results by "citizen scientists"	Desired(?) by funding agencies: people able to reproduce an analysis should be awarded … ce" – beyond what can realistically be af…
3	Reproducible results where consumer ≠ producer and outside immediate community	Stronger demonstration of lo… preservation. Knowledge stored is suff… physicist outside immediate commu… produce results
2	**Reproducible results where consumer ≠ producer but within same "larger community", e.g. LHC (ATLAS / CMS; CDF / D0, …)**	**Highly desir… minimal" long-term preser… knowledge" stored is sufficient for a p… om a different collaboration (but within … overall programme) to reproduce results**
1	Reproducible results wh… consumer = produc…	Required during lifetime of collaboration
0	N/A	Data is lost: logically or physically. This is probably the reality for the bulk of pre-DPHEP experiments (and even some of those??)

REPRODUCIBLE RESULTS AFTER "PORTING" TO NEW ENVIRONMENT!

ICFA Statement on LTDP

- *The International Committee for Future Accelerators (ICFA) supports the efforts of the Data Preservation in High Energy Physics (DPHEP) study group on long-term data preservation and welcomes its transition to an active international collaboration with a full-time project manager.* ***It encourages laboratories, institutes and experiments to review the draft DPHEP Collaboration Agreement with a view to joining by mid- to late-2013.***

- *ICFA notes the lack of effort available to pursue these activities in the short-term and the possible consequences on data preservation in the medium to long-term.* ***We further note the opportunities in this area for international collaboration with other disciplines and encourage the DPHEP Collaboration to vigorously pursue its activities****. In particular, the effort required to prepare project proposals must be prioritized, in addition to supporting on-going data preservation activities.*

- ***ICFA notes the important benefits of long-term data preservation to exploit the full scientific potential of the, often unique, datasets****. This potential includes not only future scientific publications but also educational outreach purposes, and the Open Access policies emerging from the funding agencies.*

- 15 March 2013

RDA Preservation WG

- The RDA – strongly supported by EU, NSF, AU – seen as an element of implementing HLEG 2030 vision
- A Interest Group on DP was approved in May
  - Chair: David Giaretta (APA, SCIDIP-ES, author of "Advanced DP", ex-DCC, ex-STFC)
  - Co-chair, rapporteur: JDS
- The intent is to show progress by each RDA plenary (March, September) and co-ordinate international activities, identify candidate services for standardization, lobby for funding…

RDA IG – Work steps (until **DUB**)

- **Regular virtual meetings**
  - **Contribute concepts:**
    - Use cases
    - Potential services + Relevant abstract interfaces
  - **Identify:**
    - where we can bring existing capabilities together – as proof of concept
    - "gaps" in shared preservation e-infrastructure *(to be filled via projects?)*
    - **how the work of other IGs and WGs can fit in**
    - **potential WGs arising from this IG**
  - **(Eventual) outcomes:**
    - **Preservation tool-kit, "Services", e.g. media migration**

DPHEP Component Breakdown

- Can break this down into three distinct areas
  - (OAIS reference model is somewhat more complex: this is a zeroth iteration)

- **"Archive issues"**

- **Digital Libraries & "Adding Value" to data**

- **"Knowledge retention" – the Crux of the Matter**

Archive Issues

- ✓ We (HEP) has significant experience of 100PB+ distributed data stores
- ✓ Coordination of long-term "bit preservation" issues via HEPiX
- ✓ And with other disciplines e.g. via ~~IEEE MSST~~ RDA
- × **Sustainable models for long-term multi-disciplinary data archives still to be ~~solved~~ proven**

- ➢ **H2020 funding targetted for this**

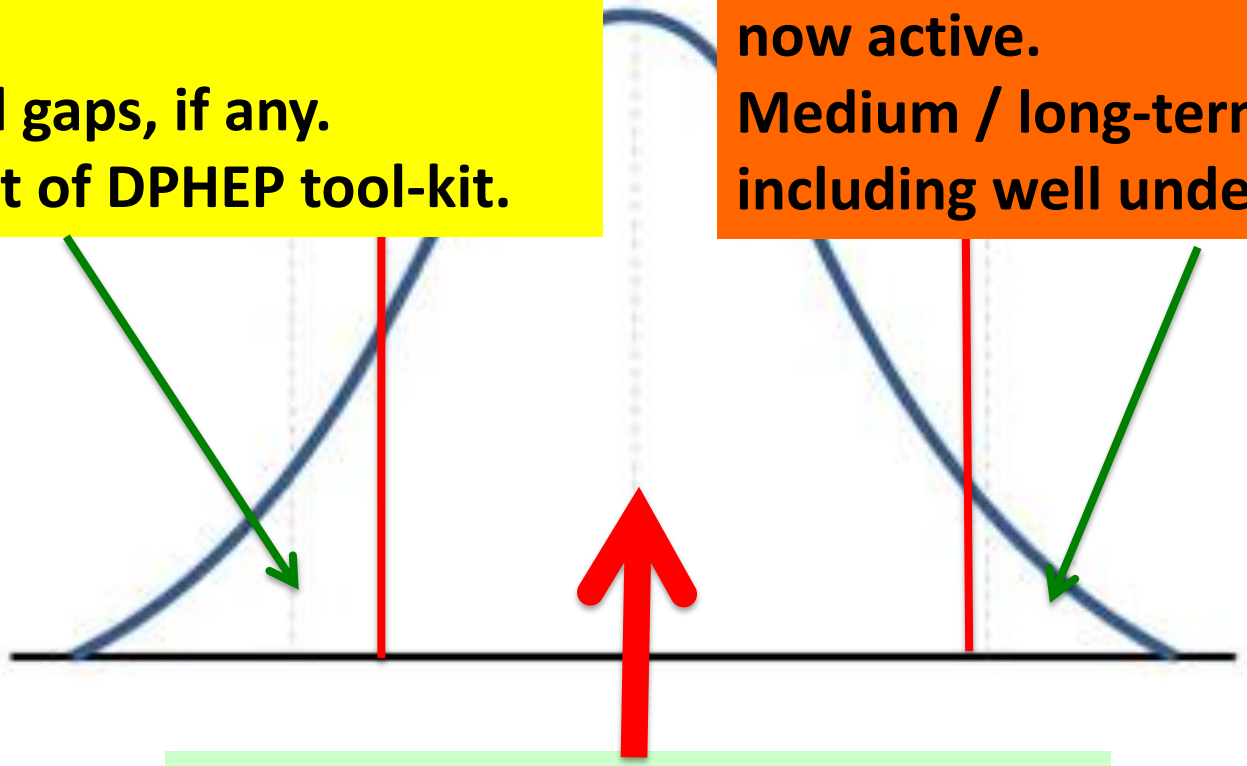
Digital Libraries

- ✓ Significant investment in this space, including multiple EU (and other) funded projects
- ✓ No reason to believe that the issues will not be solved, nor that funding models will not exist, e.g. adapted from “traditional” libraries
- ✓ Related topics: “linked data”, “adding value to data” – again with projects / communities
- ➢ **Working closely with these projects / communities, services: formalize in short-term**

From Common Projects to Services

Tools and Services, e.g. INSPIRE / Invenio.
Understand gaps, if any.
Declare part of DPHEP tool-kit.

HEPiX WG on Bit Preservation now active.
Medium / long-term plan including well understood costs



DPHEP Common Projects:
Emulation, Migration, RECAST, …

What	When
Collaboration Agreement	Q3-Q4 2013
Preparation for H2020	Now – Q3/Q4 2013
HEPiX WG in place	<Q4 2014
First H2020 calls open	Dec 2014
ICFA report (work plan, including sustainability plan)	DESY, Feb 20-21 2014
H2020 Proposal	End Q1 2014
DPHEP Portal Available	mid 2014
H2020 news	July 2014
LEP Data “recovery” (CERNLIB???)	End 2014?
Validation framework(s)	2014 / 2015?
Long-term CDI #1	2015 – 2017
Full(?) understanding of costs	**2016/17? INITIAL: SPRING 2014**
Sustainable, repeatable LTDP	201?

Summary

- **Strong links with other disciplines established / re-enforced**
- Active discussions with funding agencies + component break-down, **Use & Business Cases, Costs & Cost Models**
- **Skeleton (+ some flesh) of LT sustainable strategies: costs to be analyzed in dedicated workshop in 2014**
- **Moving rapidly towards definition and implementation of LT sustainable services, together with Common Projects**

➢ **Where do we want to be in Okinawa?**