Preservation of Data for Future Use

Jamie.Shiers@cern.ch

Update on Data Preservation Activities

WLCG CB, May 2012, NYC

Goals

- Reminder of the DPHEP session during CHEP and the status of the blueprint document;
- Information on ILDAP, DASPOS, PODDS and other projects;
- Stress the importance of doing something in common for the LHC experiments (and LEP!);
- Emphasize the of need working with other disciplines – particularly if we wish to secure any EU funding (we can also learn from them);
- Next DPHEP event & other future activities.

Partners

- Siggi Bethke
- Marcello Maggi
- Cristi Diaconu
- Ghita Rahal
- Volker Guelzow
- Tim Smith
- David South
- John Kennedy
- Giacinto Donvito
- Domenico Giordano
- Stefan Kluth
- Claudio Vuerli

DPHEP @ CHEP

- Plenary talk by David South, Wed @ 10:30
  - Data Preservation and Long Term Analysis in High Energy Physics
- Parallel session on Thursday afternoon
  - Agenda here – includes talk on future funding options
- Plenary session proposed for CHEP 2013
  - Yet more emphasis; talks from other communities & potential service providers
- Final DPHEP blueprint (V3) to be submitted Sunday, May 20, 20h00 CET

ILDAP (November 2011)

- International Long-term Data and Analysis Preservation: CERN, CNRS, DESY
  - INFRA-201203.2: *International cooperation with the USA on common e-infrastructure for scientific data* (CSA-**CA**) (coordination)
- Objectives:
  1. Through collaboration with related projects in the US, take the recommendations of the Data Preservation in HEP for Long-Term Analysis [DPHEP] study group and carry them to a wider scope scientifically and to the LHC scale in terms of volume and duration;
  2. Establish the requirements and standards to be used throughout the project and develop first a demonstrator and later a prototype based on these (see associated milestones and deliverables);
  3. Ensure the full engagement of the LHC experiments, themselves international collaborations with significant components in both Europe and the US, in data preservation activities for the data that is currently being acquired.
- Requested grant: EUR1.3M
- **Evaluation: Scientific & Technical Excellence: 2.5/5; Quality and Efficiency of Management: 3/5; Impact: 4/5; Total: 9.5/15, threshold 10/15**
- *The proposal is **narrow**, both in terms of scientific discipline (only HEP is covered) and in terms of problem domain (long term preservation), and as such is **not well aligned** with the call's objective to establish "a EU-US coordination platform aimed at full interoperability of scientific data infrastructure".*
- *The problem is **convincingly** described and the need to address it at an global scale is **well-motivated***

DASPOS

- Data And Software Preservation for Open Science (DASPOS) – Mike Hildreth
- "Sister project" to ILDAP
- A primary goal of the DASPOS effort would be to enable, toward the end of the funding period, a "Curation Challenge" where, for example, an ATLAS physicist might perform an analysis on curated and archived CMS data.

➢ **Have only seen a draft proposal and do not know the status of this project.**

PODDS (April 2012)

- PRESERVATION OF DATA for the DIGITAL SOCIETY – FP7-ICT-2011-9 (CSA-**SA**) (CERN, INFN, MPG)

- **Tried to learn from previous review!**

- This was a clean-sheet proposal that took call and expanded on it!
- *ICT-2011.4.3 Digital Preservation*
  - a) More reliable and secure preservation technologies and methods
  - b) Technologies and systems for intelligent management of preservation
  - c) Interdisciplinary research networks
  - **d) Promotion schemes for the uptake of digital preservation research outcomes** including outreach to new stakeholders and **road mapping activities**.

Funding schemes:
a) STREP; b) IP; c) NoE **d) CSA**

CERN (road mapping); INFN (digital research outcomes); MPG (outreach)

Stressed as much as possible multi-disciplinary actions

Roadmap to be endorsed by EiroFORUM members and others

PODDS Use Cases

1. *The preservation and continued ability to use e-records of all kinds for at least the lifespan of an individual (say 100 years: for some disciplines even longer retention periods are desirable);*
2. *The preservation and continued ability to re-use (e.g. re-analyse) scientific and library data for a small number of decades;*
3. *The preservation of scientific data for educational outreach for a small number of decades.*

- **Use Case 1 is fundamental to the Digital Society**

What have we learned?

1. We must work with other disciplines
2. We must work with other disciplines
3. We must work with other disciplines
4. We can **learn** from other disciplines
5. We can **help** other disciplines
6. It takes a (very) **long** time to put together an active partnership

- **The effort has not been wasted – it has raised awareness and kicked off some concrete actions that we need to continue**

Next DPHEP Event

- The next DPHEP workshop is foreseen for later this year (October?) to be hosted by the Max Planck Institute for Physics, Munich

- The intent is to involve (at least) the Max-Planck-Society and its wide range of (scientific) disciplines…
  - bio, med, phys, chem, humanities, informatics, math, astro, tech

- We have also proposed a DP plenary at EGI TF
  - September in Prague – possibility to send messages to any EU officials present? And other communities?

Other Activities

- Visibility of DP is increasing, e.g.
  - the idea to include Data Preservation in the INFN *"Piano Triennale"*
  - a project (called PREDON) in CNRS to start prospective studies on data preservation in a multi-disciplinary framework (within the framework of Massive Data Program MASTODONS)

- **European Strategy of Particle Physics – Krakow Workshop:** call for proposals
  - Propose Data Preservation as a fundamental component of future PP strategy: Siggi would be ideal candidate to do this!

DP for LHC (& LEP!)

- WLCG brings together the 4 main LHC experiments for offline physics computing
  - We also had a request to “grid-enable” ALEPH (+ other LEP experiments for good measure) for “obvious reasons” during 2011…
- Coordinated DP activities for all LHC experiments – and LEP whilst there is still time – are desirable and supported
- The coordination activities are not the manpower intensive part:
  - Active participation within the experiments needed;
  - Clear understanding of the support services needed;
  - Coordination with other disciplines highly desirable.
- For an affordable solution we will have to show some flexibility – e.g. adopt a standard format for that data that is to be preserved

- (How much of the “WLCG” + experiment stack is “standard” today)
- (Who could support such a stack for decades +)

Summary

- We need to do something **now** about Data Preservation for LHC – and LEP data!
- A number of coordination roles exist: between experiments, across labs and with others
- FP8 is an attractive funding possibility – it would be a shame to miss it!
- **Make Data Preservation an Integral Part of Particle Physics Roadmap via ESPG**
- [Don’t forget DPHEP @ CHEP]