S/W Preservation & Legacy Issues at LEP (CERN)

Jamie.Shiers@cern.ch

Former Developer / Librarian: CERNLIB

Currently: Data Preservation in HEP

Paper and much background info:

https://indico.cern.ch/event/801649/

Introduction

- CERNLIB: some 3MLOC, mainly Fortran, some Assembler(s), C(++), Pascal etc.
- Many contributors over several decades
- Source code marked up for multi-platform support (a bit like C pre-processor)
- Built for many platforms, different byte (word) length / order, FP operations, character sets
- Regular binary multi-platform releases OR
- Access to source code with build “suggestions”
- In its dying days: migrated to CVS, cpp, gmake

➢ **Now “unsupported” but still in use, still being built, still used for published papers (Oct. 2018)…**

Preservation or Deletion?

- "Obsolete" routines were first de-activated then later deleted (active) / not carried forward (passive)
- No systematic archiving of source matching each published release
- Release notes documented in Computer Newsletters (scanned; available) up to early 90s
- MANY copies distributed around the world, mainly on 1600 bpi tapes (so lost???)
- **LEP (1989 – 2000) data: ~100TB / experiment (4)**
- **[LHC (2009 – 2040?) data: getting on for 100PB (av.) per experiment]**
- Now 2 tape copies & 1 disk copy at CERN, plus numerous copies at outside labs

➢ **Would it really have been impossible to have systematically archived all s/w releases?**

➢ **N.B. at the start of LEP, no managed storage: users bought and managed their own (200MB) tapes!**

Reviving old data

- Of course, you need the data (“bits”) itself
- But this is not enough:
  - Source code; build procedures; validation suite …
- ***Still*** not enough:
  - calibrations, field-maps, basic documentation …
- In some cases, source was in PL no longer supported
  - Non-standard, e.g. CDC Fortran, Mortran, Sheltran, VAX Fortran, …
- Can you revive code when you don’t really know what it is supposed to be doing?
- With few / no comments, 1-2 character variables?

A Holistic Approach (“Space”)

- I contend that source code preservation, without build and validations systems, without basic documentation, as well as the necessary “environment” is **close to useless** for large-scale projects such as LEP, LHC, …
- The “environment” is much more than the “VM” in which the code runs **AND** includes necessary configuration files, DB snap-shots, magnetic field-maps etc.
- Even this may not be enough for reproducibility, e.g. HEP uses “triggers” (AKA filters) where much data is **DISCARDED!**

Dear Jamie:

Reading the articles on open data and re-use in the last issue of the CERN Courier has revived one of my worries which had remained latent for some time:

How do we document convincingly trigger design and performance.

The concern is not so much that what we did not select ***is lost for ever,*** *but what are the biases in the raw data we retain.*

Understanding the trigger performance is ***mandatory*** *if outsiders, by data re-use, produce important discoveries exposed to scrutiny by the scientific community.*

The LHC experiments are making a laudable effort to make their data available but I have the impression the issue of documentation / knowledge is not addressed vigorously enough at its roots, i. e. real time trigger.

I feel that validity and limitations of data re-use for discovery should be the subject of a thorough and humble analysis.

I am available for coffee if you wish to discuss this issue.

Cheers, Pier Giorgio

A Holistic Approach – Time

- OAIS tells us that we have to be aware of – and plan for – changes in technology
- But this applies also to services, to programming languages, operating systems, compilers, code repositories etc.
- It will also apply to PID systems etc!
- Real experience with migrations show that they are expensive, often uncover hidden bugs, often data / knowledge is lost in the process
- And this is during the “active” phase of a project!

Holistic in Space & Time

- Time – addressing service / technology / personnel changes and associated migration, adaption, re-validation, …

- Space – addressing s/w and its complete eco-system, including also the data for which it is intended / associated