Long-Term Data and Software Preservation

Pere Mato/CERN

2/12/2013

LTDP Challenges

- "Bit preservation" is probably not very difficult
  - You need to ensure that data is migrated from a media/format to a more recent media/format at regular basis
  - Easily findable, fully usable by designated communities with clear access policies
- Data interpretation and knowledge preservation is much more difficult
  - HEP data is, in general, not easy to understand by people outside the immediate community, specially true for unprocessed data (e.g. RAW, DST, ...)
    - Details of the detector, details of the experimental conditions, details of the calibrations, details of ...
  - The **experimental software** is essential for the interpretation of data

Knowledge and Software needs "also" to be preserved

Software Preservation Approaches

- Porting Software
  - Keep software alive by porting from system to system, compiler to compiler, etc.
  - Difficult and very expensive in manpower (e.g. migrating CERNLIB to 64 bit, ROOT 3 to ROOT 6, etc.)
  - Validation, validation, validation
- Freezing the Software
  - Virtualization together with network isolation can really help here
    - Virtualization still too young to know how it will “age”
  - Controlled experiments are needed
    - Take very old (~20 years) experiment software (plus compiler, OS, etc.) and resurrect it to perform a number of ‘use cases’
    - Learn from the encountered problems
    - Provide guidelines to new and running experiments