technische universität
dortmund

Discussion on Data Preservation Models

David South (TU Dortmund) and Homer Neal (SLAC)

Second Workshop on Data Preservation and Long Term Analysis

SLAC
NATIONAL ACCELERATOR LABORATORY

26 - 28 May 2009

First Steps

- Started by thinking about the different potential levels of data to preserve, and different possible uses with each of these levels
  - Consider each level of complexity separately
  - In the more complex scenarios, each type of project (ee, ep, pp) should be thought of separately, in fact some things will stay experiment specific
- Aim of the session: Arrive at a structure for this section, including some "definite statements"
- There is some overlap with the "Physics Cases" here, which can drive what you want to preserve, but we try to stick to the models themselves here..

Defining the Different Data Levels

- Found this was best splits in terms levels of complexity of the levels: Sub-sections of the draft

Increasing complexity

(inclusive)

  - Nothing at all
  - Additional documentation for existing results
  - Data only *(incl. Simplified Common Data Format)*
  - Data + "community code" (ROOT, ..)
  - Data + experiment specific software (SIM+REC)

- Include a description of each of these cases, including:
  - Effort needed (two stages: preparation, maintenance)
  - Estimated costs in FTE years
  - Benefits of preservation

May be different depending on which stage the experiment is at

It’s not about Data Volume

- Perhaps stating the obvious, but:

“In most cases the (even raw) data volume will not be significant with respect to available storage media size, and therefore the preservation model - and the associated costs, person-power demands and so on - should reflect what the aim of preservation is: what do we want to use the data for?”

1. Doing Nothing At All

- This option should be included, although of course *with care* and be presented in terms of what the consequences of doing nothing at all are
  - Costs: 0
  - Benefits (of NOT doing this!): References to previous physics case success from past experiences: lessons learned

2. Additional (better) Documentation

- A model of preservation, without actually preserving the data, is to provide additional documentation: *which is also a required effort of all subsequent preservation models*
- Additional documentation to include:
  - More information associated with publications
  - Internal collaboration notes
  - Technical drawings, general info about the experiment
  - Various forms of metadata
  - General experimental studies (systematic correlations, ..)
  - Hypernews
- Costs: 0, runs subsequent
- Benefits: self-explanatory

2. Additional (better) Documentation

- Recommendations:

“Experiments should consult with an archivist early on.”

“Documentation [including those forms listed on the previous page] should be coherently stored in a system like INSPIRE. This can be password locked for the lifetime of the collaboration, then made public later.”

“A wiki-like system should be employed for day to day documentation purposes, ie for instructions on how to do analysis. The information contained here, and in any associated hyper-news, can be then ported to INSPIRE at the preservation stage.”

2. Additional (better) Documentation

- Recommendations (continued):

"Auto-documentation tools like those included in ROOT should be used to their maximum ability."

"A common format for popular tools (e.g. electronic log books) would be useful, enabling such metadata to be preserved in a similar way."

3. "Just the Data": 4-vector format

- The simplest form of data preservation: reduce the data to simple 4-vector format
- Does not need reconstruction / simulation software
- Candidate for common (combined?) data format
- Perfect for Outreach projects, using HEP data as a teaching tool and to attract new interest in the field
- We can learn from previous attempts about the *unsuitability* of such formats for "real" (publishable) analysis (QUERO..)

- Costs: small num. of FTEs, minimal w.r.t. more complex schemes
- Benefits: Relatively easy to prepare, clear value of Outreach

3. "Just the Data": 4-vector format

- Recommendations:

"If the experimental data is to be preserved, it should be done in a simple structure as possible."

"The data should be sufficiently self-describing."

"A simple four-vector format can be very useful in terms of a model for Outreach purposes."

"A common data set between [similar] experiments should be possible in this format."

"However, it should be clear that this format will not be sufficient to perform a full physics analysis."

4. Data + Community Code

- Adds an additional layer of complication to the data format, but benefits e.g. ROOT features and improvements
- Relies on longevity of the community [analysis] code - first set of software which also needs preserving
- Relies on longevity of external software, again e.g. ROOT

- Costs: couple of FTEs; crucial impact in preparation stage of streamlining of analysis level software
- Benefits: Ease of analysis

5. Data + Experiment Specific Software

- In order to preserve the ability to do a full, scientific analysis the reconstruction / simulation code is needed
- This may or may not require RAW data, depending on what is stored on the DST level - which is experiment specific - but generally for greater flexibility all data should be preserved
- Various Physics Cases outline this need
- Attempting to write a new simulation towards the end of the collaboration is not so useful: requires much verification

- Cost: significant number of FTEs for preparation phase, fewer in the maintenance phase
- Benefits: Reproducible full physics analysis chain and full flexibility for future use

5. Data + Experiment Specific Software

- Recommendations:

"At this level of preservation the aim is not for a common format but rather a common standard."

"Here is more a discussion of *analysis preservation* rather than data preservation."

"In order to perform [new] MC simulations and to derive associated corrections, studies of efficiencies and acceptances, and a full systematic error analysis the full data and analysis chain should be preserved."

"So called 'FAST' simulations may be useful during the lifetime of the experiment but the full simulation should be the version that is preserved."

5. Data + Experiment Specific Software

- Recommendations (continued):

"To maximise the efficiency of such a large preservation project, a collaboration should employ as much centralised software as possible. This also benefits the collaboration: greater adoption of community code, results in a more efficient use of person-power."

Final (personal) remarks

- I think from our discussions there is both room and a desire for both an outreach format and preservation of the full analysis chain

- The 4-vector, “just data” model may well be useful for Outreach purposes, but I think it should be stated that that is what it is for

- In reality, only an ex-collaboration member will be able to perform a full analysis, given the complexity of what we do, and in describing such a preservation model it should also be made clear that the intention is not to enable this to be done by anyone

Not Covered, Additional Points

- Type of preservation model:
  - Static, frozen: Virtualisation techniques
  - Updating OS, ROOT (dynamic “rolling” model)
- Link to technologies
  - How will the preservation models benefit from advancements in technology
  - What known faults of technology must the preservation models be protected against
  - Make suggestions for hardware and software technologies that should be pursued