Scenarios for long term analysis (Summary)

Stephen Wolbers

Fermilab

Workshop on Data Preservation and Long Term Analysis in HEP

DESY, January 26-28, 2009

Outline

- General Comments
- Scenarios for long-term analysis
  - Experiment (completed, running, future)
  - Joint experiment (ee, ep, pp)
  - HEP wide
  - public
- Custodianship issues
  - Code, environment, expertise, etc.
  - Data Center Considerations
- Examples to learn from
  - Astronomy
  - Other science data preservation

Some comments

- I am new to this discussion, hence the slightly unforcussed nature of this summary
- To a large extent experiments have analyzed data for as long as they had the effort and expertise to do so.
- Once the effort and expertise was "below threshold" then outside factors had an impact:
  - Tapes were thrown away or lost or unreadable.
  - Code was lost (kept in home areas, backup tapes were lost, etc.)
  - Documentation was incomplete or was only known by experts
  - Operating systems changed and code broke
  - Data handling mechanisms stopped working
- By and large these barriers were overcome by an experiment with strong motivation to continue analysis
  - This was usually done on a case-by-case basis

General considerations (1)

- Are we too late?
  - For many experiments (thinking of a much broader selection than the subset considered here): yes
    - Old bubble chamber film – gone
    - Old tapes – gone or unreadable for a large number of experiments
    - Software and expertise – long gone
  - For the experiments considered at this workshop:
    - Worth investigating and understanding
    - Bring the entire community up to speed
      - Not just the experts who are thinking about this issue
  - A proposed split:
    - Completed experiments
    - Running experiments
    - Future experiments

General considerations (2)

- **Why preserve data and analysis capability?**
  - Physics case must be made:
    - Necessary but not sufficient
    - Wishful thinking is not good enough
    - May require a real shift of resources or a choice between this and some other activity
      - Not just for scientists but also for computing professionals and computing resources
- **Why hasn't it been done in the past?**
  - Probably doesn't matter but it would be nice to understand as people think about future options
- **What are the use cases?**
  - Specific analyses
  - Speculative or other analysis of "unique" data
  - Follow-up or contribution to future discoveries or measurements
  - In the end these need to be defined to make progress

Completed Experiments

- Need to define the requirements or goals
- Define more clearly what is required for data preservation
  - Starting with the definition and goals of the term "data preservation"
- Many issues:
  - Data samples need to be defined
  - Code and calibration and related information
  - Documentation!
  - Expertise
  - Timescale for analysis (few years, forever?)
    - Time until a follow-on experiment
  - Effort estimates (very important)
  - Different experiments may have different requirements, these should be documented

Completed Experiments (2)

- Authorship and collaboration ownership of the data
  - Experiments typically run with some sort of collaborative agreement and management board
  - The issue of how to handle analysis results coming from these experiments must be defined and approved
  - Define the "end" of the collaboration responsibility and restrictions
    - Fixed time-frame
    - Tied to data or physics topics
    - Forever
- We heard some ideas from HERA, LEP, others
  - Need to expand to other interested experiments
  - The criteria for who to include needs to be worked out

Running Experiments

- Mainly focus on Tevatron at this meeting
- Same issues as completed experiments except:
  - Expertise is by and large still in place
  - Effort is (potentially) available to prepare for a number of possibilities:
    - Analysis for the collaboration for a targeted number of years
    - Migration of the analysis and data for longer period of time
    - Ntuple preparation for potentially very long-term analysis
    - Preparation of standard format of data
    - Combined analysis
- Include other experiments as appropriate

Documentation is not just information about the data and data analysis

- Documentation
  - The experimental apparatus
  - The beam and beam conditions
  - Experiment performance and issues important for physics analysis
  - Many internal notes – how are they going to be kept?
  - Information on user's home directories
    - Wiped out when people leave in many cases
    - Backup tapes are short-lived
- Place to store the information
  - INSPIRE, the equivalent, national/international repositories
    - Need to decide
  - Can provide location and links for much of the above

Future Experiments

- Likely to be the most important case, or at least the one with the greatest opportunity to do something planned and new
- One can imagine the proposal, funding and approval process integrating the long-term data access and preservation component
  - Best time to organize the thinking
  - Make commitments
  - Acquire funding and effort
  - Make decisions that are consistent with the requirements
  - Need a threshold for the process:
    - All experiments
    - Large experiments
- Opportunity to lay out a plan

LHC and others

- LHC experiments are not discussed at this time
- Hard to ignore such a huge community and data sample, at least in planning
- Many other experiments are taking data or will take data soon
- It is understood that taking on too much at once may diffuse effort and cause this process to diverge
  - But the question will likely come up
  - Should be addressed

HEP-wide Access

- Does it make sense?
- Physics is invariant
- Combinations of experiments are critical for the best physics
  - Done in a pretty ad-hoc way
  - HFAG, CTEQ, LEP working groups, etc.
- Some ideas were heard
  - Quearo or some common data representation
  - Ntuples
  - INSPIRE and links available
  - Other

Joint or combined analysis

- Important consideration
- Need to understand how much is required and beneficial
- Make commitments for access to:
  - Distributions
  - Ntuples
  - Event-by-event data
  - Common data formats

Federation of data

- Can HEP make available detailed information from many experiments?
- Requirements:
  - Common format (is this possible?)
    - Storage standards
    - Object standards
  - Place to store data
  - Mechanism for serving data
  - Funding mechanism
  - Support
  - Standards for interoperability
  - Management and oversight

Public Access

- Interesting and important issue
- What does it mean? Need to define.
  - Ntuples
  - Web interface with tools
  - Standard data format
  - Each experiment individually
  - All experiments
- Publishable results?
  - Publication process
  - No internal review
  - Referee?
- Authorship and responsibilities

Custodianship

- Responsibility for code, associated information needed to analyze, and the expertise needs to be defined in some way.
  - Could be open source
  - Kept in a data center
  - INSPIRE – protected by access rights
  - Other
- Expertise
  - Collaboration-owned?
  - Released to world (documentation)
  - Funded by agencies in a formal way
  - Other

Data Custodianship

- Data centers or the equivalent can be defined to be data custodians
- Long-term data custodian
  - Data preservation agreement with the experiments and other internal and external entities
  - Disaster planning and recovery
  - Duplicate data when necessary
  - Data migration to higher density media automatically
  - This should all be spelled out and not left to chance
- Long-term data access issues
  - Data format
  - File structure
  - Methods for access to data
- Budget – taking care of data is not free!

Examples to learn from

- Astronomy
- Data Archive is used for analysis and for public access
- Formal standard FITS format (1977)
- Planning from the start to make data available
- International projects to keep the data
- Reprocessing is done just after the data is taken
- More should be done to understand what was done and how it might apply
- Not perfect but interesting

Other science preservation

- Clearly people are interested in maximizing the physics capability of the experiments
- Funding agencies and governments would like to ensure that information is not lost
- We need to follow what is happening and participate as appropriate

Conclusion

- Everyone agrees that data and analysis preservation is desirable
- Need to understand what is involved in many areas
- Not free, in some cases may not even be possible
- The press of new efforts and budget cuts tend to push this activity away
- Suitability for this work in HEP (students, postdocs, others) has to be understood
- It was a stimulating discussion and an enjoyable workshop!