FERMILAB-CONF-09-073-CD

Scenarios for Long-Term Analysis
Proceedings of Data Preservation and Long-Term Analysis Workshop
DESY 26-28 January 2009

Stephen Wolbers
Fermilab*

Abstract
Data Preservation and Long-Term Analysis of High Energy Physics (HEP) Experiments data is described and summarized in this talk. The summary covers information presented at the First Workshop on Data Preservation and Long-Term Analysis. Experiments representing e^+e^- collisions (LEP, B Factories and CLEO), ep collisions (H1 and ZEUS), $p\overline{p}$ collisions (CDF and D0) and others presented interesting information related to utilizing the large datasets collected over many years at these HEP facilities. Many questions and issues remain to be explored.

1 General Comments and Introduction

1.1 Introduction

High Energy Physics experiments historically have analyzed the data collected by each experiment for as long as the collaboration has the effort and expertise to do so. Once the effort and expertise available to the experiment fell below some threshold then other factors came into play. These factors include the following:

- Data tapes are thrown away or are unreadable (for many reasons).
- Software and calibrations and/or other necessary information is lost.
- Documentation is incomplete or only understood or understandable by experts.
- Operating systems and compilers change. Software is no longer usable without a porting effort.
- Data handling mechanisms stop working.
- Necessary tapedrive technology is not available.

1.2 General Comments

One obvious question to ask about the issue of long-term data preservation and analysis of data is whether it is already too late to make an effective plan that enables effective use of the data. Many examples from the past show that the ability to perform long-term analysis is severely hampered or is not possible because data, software and expertise slowly or sometimes rapidly disappears

*Operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy

over time. For example, bubble chamber film is all gone by now, old data tapes are thrown out or are unreadable and the software is often lost when backup tapes are stored, unreadable, thrown out or overwritten. Experts move on to other experiments or out of the field. Completed experiments with no intention or plan for long-term analyses are by default making a decision not to do so.

Those experiments that have recently completed or will complete in the near future are the most likely to succeed and are worth full investigation at this time. In addition, it makes sense to bring the entire community up to speed on this issue.

A clear physics case must be made to justify the money, time and effort that will be required for successful long-term analysis of the data. Though necessary, this is in no way sufficient for a successful program. In the end there may be a real shift in resources from current and future experiments to the preservation activity and this must be considered as part of the case that must be made.

2 Data Preservation and Analysis Issues and Considerations

2.1 Completed Experiments

For completed experiments there are a list of issues that should be addressed when making a case for the resources required for long-term analysis of the data. This includes the goals, the requirements and the definition of terms such as "data preservation". The data samples must be defined, code and calibration and related information must be located and collected and the importance of each must be understood, documentation must be located and/or written, expertise for building and managing the system or systems must be clearly defined and acquired and maintained, timescales for analysis must be made clear, effort estimates need to be made and reviewed, and experiment-dependent issues must be understood and documented.

A very important issue, mentioned often the workshop, is one of authorship and credit for scientific findings resulting from analysis of the data. HEP has certainly learned that even in running experiment with an active collaboration the issue is one that requires attention and policies vary from experiment to experiment. This issue will become even more important for the long-term analysis of the data, once the collaboration becomes more or less "inactive". Most large HEP experiments handle this issue and document it in a collaborative agreement to handle questions or problems that can arise during the physics approval and publication process. However, in the case of an inactive collaboration a clear understanding of process and rules will need to be defined. We heard some ideas at this workshop but this does need to be expanded for further elaboration and understanding.

2.2 Running Experiments

At this workshop the primary running experiments heard from were the Run 2 experiments CDF and D0. In general these experiments have and will have exactly the same issues to deal with as they wind down and focus attention on the long-term analysis of the huge data samples that have been collected. They have the advantages that the expertise is by and large still in place (albeit shrinking) and that effort is potentially available from within the collaboration to prepare for long-term analysis.

2.3 Documentation

An interesting issue for all experiments and collaborations is documentation. Many aspects of an experiment require proper documentation, both for the experiment as it is constructed and operated, but also as part of the proper analysis of the data. This includes documentation of the experimental apparatus, the beam and beam conditions, the experiment's performance and how that relates to the physics analysis. This is usually, but not always, contained in voluminous internal notes, maintained by the collaboration. In some cases the information is kept in personal notebooks, electronic logs, code kept on individual computer accounts, or other places not properly maintained for long-term access. An interesting initiative that may provide a mechanism for this information, INSPIRE [1], was described at this workshop.

2.4 Future Experiments

The greatest opportunity to properly plan for long-term data preservation and analysis is with future experiments. One could imagine that new experiments can be required or at least requested to provide in their proposals and funding requests a plan for long-term analysis and data preservation. This would naturally include the funding, effort, and organization required to accomplish this as an integrated part of the experiment rather than an after the fact activity.

2.5 HEP-wide access to data and combined analysis

Many collaborations and HEP scientists are thinking about wider access to data for the purposes of performing combinations of physics results or data. Many such combinations are performed today in ad-hoc but fairly well-established ways by HFAG [2], CTEQ [3], the LEP working groups [4], the Tevatron working groups [5], HERA working groups [6], PDG [7], etc. Ideas for moving forward include creating a common data format, possibly something like the Quaero [8] approach, making NTuples available, or using INSPIRE as a common framework. The broader question of creating a common format of event quantities is an interesting issue to study. There are many questions about how this could be approached and how to engage sufficient resources to study the idea.

2.6 Public Access

Many talks at the workshop discussed possible public access to HEP data. It is an important and interesting issue. HEP research is funded by the public and there are good reasons for the public to expect that the researchers make effective use of the data collected by the experiment. Whether that includes public access to the data in some way is something that should be explored. As in previous sections the question of just how the data would be formatted, accessed and documented is not so clear. The question of how scientific results might be published is another area of concern. The analysis would not be subject to rigorous internal collaboration review. Authorship would have to be thought about as well.

3 Custodianship

3.1 Data

In most cases today data centers are the custodians of the large data sets from HEP experiments. This is usually satisfactory both for the duration of the experiment and for some number of years after the experiment is complete. However, it is not satisfactory in terms of making a plan for long-term analysis. A custodian or custodians for long-term data storage must be determined and agreements made for all aspects of this task, including data migration, disaster planning and possible data recovery and duplication, and funding for the activity. This should all be spelled out and not left to chance. Issues such as data format, file structure, and methods of accessing the data all need to be included in the agreements.

3.2 Other resources

There are many other areas of responsibility for the long-term preservation of code, calibrations, documents, expertise and other necessary resources for the proper analysis of data. These will have to be defined if the custodianship is to be properly established, funded and maintained.

4 Examples from related fields

HEP is not the only field studying long-term data storage, analysis and access issues. A well-known example comes from Astronomy and Astrophysics. In this area data archives are often used for analysis and for public access to the data. A formal standard format (FITS [9]) was developed in 1977 and is used to store the data. Projects and experiments plan on making the data available as an integral part of the experiment. The data is released to others outside of the collaborations and to the public as part of well-defined data releases. Though not a perfect system there are some ideas here that HEP can benefit from.

It is clear that people are very interested in maximizing the physics capability of the experiments. This is true no matter how large or small the experiment is or how unique the data sample might be. Funding agencies are also interested in the data and information from the experiments and are quite concerned that the information is preserved. HEP will have to follow what is happening and be sure to participate as appropriate.

5 Summary

The HEP community agrees that data and data analysis capabilities should be maintained well beyond the end of the data-taking of the experiment and beyond the end of the formal existence of the collaborations. This is not a trivial capability to provide and the physics case and details of implementation need to be understood in some detail. It is certainly not free and in some cases may not even be possible. In HEP, as in any field, the funding and priorities for new experiments and projects will tend to reduce available funding and effort for data preservation and long-term analysis. Decisions need to be made if there is a choice between doing something new (a new accelerator, experiment, R&D project) and maintaining long-term data and analysis capability. There was little discussion of the suitability of this work for students, postdocs, computer professionals, or others who might be needed to implement and maintain the necessary systems.

The discussions and talks were stimulating and will go far to preparing for a summary and proposals for long-term data preservation and data analysis in high energy physics.

References

[1] R. D. Heuer *et al.*, arXiv:0805.2789 (2008).

[2] E. Barberio *et al.*, arXiv:0808.1297 (2008).

[3] CTEQ, http://www.phys.psu.edu/ cteq/.

[4] See for example, http://lephiggs.web.cern.ch/LEPHIGGS/www/.

[5] See for example, http://tevewwg.fnal.gov/.

[6] See for example, http://www-zeus.desy.de/physics/sfew/PUBLIC/sfew_results/preliminary/moriond04/zeush1.php.

[7] C. Amsler *et al.*, Physics Letters **B667**, 1 (2008).

[8] V. M. Abazov *et al.*, Physical Review Letters **87** (2001).

[9] See http://heasarc.gsfc.nasa.gov/docs/heasarc/fits.html.