ARCHIVING OF PARTICLE PHYSICS DATA AND RESULTS FOR LONG-TERM ACCESS AND USE

J. YEOMANS*

Scientific Information Service, CERN, Geneva, Switzerland
**E-mail: joanne.yeomans@cern.ch*

Preprints and published material are not the only output from high energy physics research that should be archived for future generations. Data are frequently not stored long-term and yet examples have arisen where such storage has been proved necessary. There are also lost possibilities for training, and indeed the records of science for future generations are diminished by the absence. Lessons learned from previous attempts and from other fields for which experimental data are successfully stored, can be used to build a storage paradigm for the future. Data from particle physics experiments are highly complex but a collaborative effort from IT staff, librarians, and physicists can perhaps have success. Issues requiring consideration include: who will have the right to access the data; how will access rights be managed; what level of data should be stored; for what length of time should the data be stored, and what additional information associated with the data must be collected. Technical problems associated with the storage and future use of analysis software must also be tackled.

Keywords: experimental data; open data; archiving.

1. Lost Data

For nearly fifteen years electronic physics preprints have been submitted to repositories, stored, indexed, retrieved, and shared, to the benefit of the high energy physics community; however, every year data from the experiments of that same community are to its detriment discarded, lost and forgotten. The very progress in technical capability that provides better means to remedy the situation, also leads to the generation of far more complex data which in turn complicates the problem. The expense and rapid advancement of new experiments makes it essential that previous results remain accessible for accountability, re-analysis, and training of future generations.

2. Preprint Management as a Successful Blueprint

2.1. *Existing Repositories*

The Cornell-hosted database, arXiv.org, is the most famous example of a preprint repository in the world even outside the physics community it serves. It is partnered by other high energy physics repositories that perform slightly different functions and are managed in the libraries of other institutes around the world, for example the SLAC SPIRES[a] group of databases and the CERN Document Server (CDS)[b] which concentrates, though not exclusively, on the institutional output of CERN itself. SPIRES and CDS both use harvesting techniques to pull records from the arXiv database. Estimates suggest that these databases host preprints or postprints of between 70-100% of the published literature. Some of these preprints are submitted direct to the repositories by the authors themselves but a significant number are retrieved from other sources by library staff. The result of these efforts is that not only can readers access published work irrespective of whether they have the necessary journal subscriptions, but that readers can locate and read those documents in a single location with which they can familiarise themselves and over which they can have some influence. The libraries concerned have a large amount of expertise and experience in managing these repositories and are increasingly working at an

[a] http://www.slac.stanford.edu/spires/
[b] http://cdsweb.cern.ch/

international level to make efficiencies between the different databases.

The set of physics repositories have come to be regarded as successful models for a world that has become interested in open access to research output and research knowledge management. Enhancements to these services which are already available or imminent will provide for an even more elaborate environment where users (readers and authors) can navigate between documents using semantic links, store details of documents as a personal subset, mark, annotate and share documents easily with colleagues, and re-use details from documents more easily in the creation of new works. In short the existing physics repositories are becoming even more of a focal point for the information needs of working physicists.

2.2. *Expansion of Repositories*

At the moment, repositories are primarily designed for readers but there is technically nothing to stop the introduction of features which could aid the authoring process. Tools to manage these things already exist and there are both financial and ergonomic arguments for bringing the pre-publication and post-publication processes more closely together.

There are also good reasons to consider improved links between the papers and the data associated with them. The presentation of data in published form is still unnecessarily governed by the limitations set by the old, printed paper era. Electronic files can now easily be attached to the main body of the work and can contain datasets related to the work described. Alternatively, there exists the possibility to link from documents to datasets hosted elsewhere.

3. Defining Data Needs

Three main groups outside the experimental collaboration itself could gain advantage from increased access to data: future researchers (who might need to re-analyse the data); contemporary researchers (who wish to reinterpret the data), and students (who can be trained using the data).

As well as a decision about whether any of these groups would be entitled to such access, there are further questions to consider:

- What tools are needed to make use of the data?
- What care must be taken to ensure the interpretation of the data is accurate and understood?
- How much of the data is useful?
- Who can access the data and when?
- How long does the data need to be archived?
- How should the data be cited?

These are not easy questions to answer and the technical solutions may not yet exist. However, some past experiences and projects in HEP and in other fields can be a source of some expertise.

3.1. *Example: LEP Data*

In 2001 the CERN IT Division and the LEP Experiments agreed that access to data would be required until at least 2006[1] and even after that date, there existed a possibility that LHC results would prompt a need for re-analysis. In order to prepare for such a situation, a plan was proposed which required the data to be stored on CASTOR (CERN Advanced STORage manager), the storage system devised for the LHC experiments, and required the latest version of some of the analysis software to be preserved on a “museum system” which could be accessed by any authorized person. Confronted with time pressure for completing the physics analyses with decreasing resources, each LEP experiment worked on its own policy, the results of which tended to focus on access rights leaving the technical solutions only vaguely formed; the level of success has therefore been mixed. An interim solution was adopted by leaving the analysis software on isolated PCs. Most of these have not been

migrated to the most recent Linux releases which are incompatible with some features of the analysis software.

There were important lessons learned from this exercise. The problems with software storage were found to be not trivial and there were also difficult decisions about the level of data that required storage. Depending on the decisions taken, data could become little usable or little useful. Without a large amount of effort in encoding certain types of information, there was also a large reliance on human memories. This and other set-backs showed that the problems were not entirely technical and a certain amount of thought was needed to define access rights and long term solutions. The example shows that pro-active steps for data archiving should be taken well before the winding down of large and complex experiments.

3.2. *Example: HEPdata*

For several decades the HEPdata service[c] has been managed by the Durham Database Group based in the Centre for Particle Theory at Durham University in the UK. The Reactions Database contains the exact numerical results behind the plots contained in published articles – any graph so included can therefore be recreated accurately by taking the values of the points. Without this extra detail, readers would have to take a pencil and ruler to try to estimate the points on a graph. It is a simple, but valuable, enhancement to the contents of the published article.

One option for improving the accessibility of data from high energy physics experiments would be to expand the coverage and the depth of data made available in this way.

[c] http://www-spires.dur.ac.uk/hepdata/

3.3. *Example: Astronomy, Chemistry and Biology Data*

There are many examples of data storage success in the astronomy, chemistry and biology fields. The National Space Science Data Center manages the archive of NASA's observational data from its space missions. The data is generally made available to the public within six months and in some cases the intention is to archive the dataset indefinitely. The CombeChem and eBank UK projects are working on data storage and re-use in chemistry and crystallography. The latter project is exploring the use of links from datasets to publications and e-learning. The European Bioinformatics Institute which is part of the European Molecular Biology Laboratory (EMBL) manages a number of biology databases to which researchers can submit data and make use of linked analysis tools.

Future high energy physics data storage projects should seek lessons and advice from the staff working on these successful databases. For example, advice on using standardized metadata definitions could benefit the building of future data relationships between the fields. Although the datasets for astronomy, chemistry, and biology tend to be less complex than those produced by typical experiments at particle accelerators, some issues are similar, such as questions of access, metadata definitions, and methods of software storage.

4. Future Collaboration for High Energy Physics Data Storage

There are many good reasons for storing particle physics data. The LHC will run for such a long time and generate such large volumes of data that it would be better to have solutions in place before the experiments come to an end and human knowledge is lost.

The obvious people to work on such a problem are the physicists themselves and the IT staff who will be involved in the data

storage. What has not been recognised during previous data storage projects in this field is the role that librarians can play in this work. Librarians have expert knowledge on metadata description and long-term archiving solutions and can provide a link with libraries in other fields who are successfully working in this area. In particular there is sense in working in line with standards for data storage in the astronomy field where there could be advantages in building bridges between the datasets in future.

A simple yet beneficial first step would be to expand the contents of preprint repositories by including extra data related directly to the contents of the preprints contained there.

Acknowledgments

Hans Hoffman has been interested in these ideas and encouraged contacts between the author and a number of people who have experience or interest in data storage. The conference presentation was prepared after discussions about data storage with many people around CERN including contributions from Luigi Rolandi (CMS and Aleph), Salvatore Mele (CMS and L3), and Ian Willers, Martti Pimia, Dirk Samyn and Zhechka Toteva from the CMS Information System iCMS.

References

1. Andreas Pfeiffer, LEP Data Archive. Last edited Apr 08 2004. URL: http://pfeiffer.home.cern.ch/pfeiffer/LEP-Data-Archive/Scenarios.html