HERA Data Preservation Projects

October 2010

The DESY Data Preservation Group

An overview of the plans and activities of the DESY Data Preservation Group was submitted to the PRC in April 2010, detailing various potential preservation projects[1]. The pupose of this document is to provide an update on those activities and to provide manpower estimates for the proposed projects at DESY. Since April, those involved have continued to develop ideas and to identify future working directions. The activities of the group were presented at the fourth DPHEP workshop[2], which was held at KEK in July 2010, where the efforts at DESY were well received by those representing international HEP community.

Following on from the initial recommendations of the DPHEP group[3], a 'Blueprint for Data Preservation in High Energy Physics' is to be published in 2010, and will include manpower requirements for data preservation in HEP. As well as reporting the neccessity of a centralised DPHEP project chair, manpower estimates for projects at the experiment, lab and international level are also to be included. The following text describes the requirements[4] for data preservation projects to ensure a long term HERA data facility is secured at DESY.

Analysis Software Validation Project

One of the main proposed data preservation projects at DESY is the development of an analysis software validation framework. Such a framework, which allows a rigorous test of experiment level software builds against changes in operating system and/or external software, is realised using virtualisation techniques and will prove invaluable in dealing with future migrations. A mock-up version has now been successfully installed, where the stability of a variety of software from the H1, ZEUS and HERA-B collaborations is tested against three different operating systems, showing the proof of principle of such a scheme. Data analysis on virtual machines has been tested by the HERMES collaboration, who could also participate in this project. The current status was presented at the DPHEP workshop at KEK, where other experiments showed interest in the further development of the project, including BaBar, who are also investigating such a validation system. The full version of the validation framework will require an injection of financial support. A position of **1 FTE for 1 year** within the DESY-IT division for the initial development and impelementation is required, followed by about **0.5 FTE per year** for the maintenance and running of the framework. The experimental contribution has been limited to small test examples so far, but the implementation of a full validation scheme of the experimental software to interface the framework developed by IT will require **1 FTE for 1 year** per participating experiment for the initial phase, followed by around **0.5 FTE per year** to provide the necessary support from the experimental side.

[1] "HERA Data Preservation Plans and Activities", DESY Data Preservation Group, submitted to PRC 69

[2] 4th DPHEP Workshop: http://indico.cern.ch/conferenceDisplay.py?confId=95512

[3] "Data Preservation in High Energy Physics", DPHEP Study Group, arXiv/0912.0255

[4] Note that standard, day-to-day computing and software activites within both the IT group and the experiments at DESY are not included in these requirements, and are assumed to be covered by other resources.

An Archival System for HERA Data

The scope of the validation framework described in the previous section does not foresee an examination of the condition of complete data sets, but rather the use of smaller samples to test software changes. As previously described, the present dCache storage system at DESY-IT is not suitable for long term storage of HERA data. Additional complications arise due to the widespread use of HEP specific protocols. It is therefore proposed to develop a long term archive system, which would include: automatic migration to new media generation and technology, automatic data integrity checks, retrieval of the data themselves and metadata operations. Initial studies are currently being carried out, and an estimated **1 FTE for 1 year** is required within DESY-IT to develop an archival system for long term, reliable storage of the HERA data. Dedicated manpower should also be foreseen for the maintenance phase. Possible links with similar needs in photon science are being investigated.

Using HERA Data for Educational Purposes

Outreach and education using a simplified format of actual HERA data is very attractive, and would be seen favourably in the HEP community. Such an initiative is complementary to efforts in high energy physics to further the public understanding of science, such as *Teilchenwelt*[5], *QuarkNet*[6] at Fermilab and *KworkQuark*[7] at DESY. Furthermore, if a global effort could be employed, via DPHEP, to produce a coherent inititaive involving data from many experiments, the results could be more rewarding. Within DPHEP, such projects already exist, such as those begun at BaBar[8], which have produced simple yet effective examples of expanded access to HEP data. The presentation of the BaBar project is in the form of a *mediaWiki*, meaning an expansion of the current offering would be possible as a collaboration between the HERA experiments and DESY-IT, who already support this flavour of wiki. Such a project can only be successful if the HERA data themselves are backed up with well thought out, pedagogical explanations and exercises, and will therefore require a significant effort. Around **1 FTE for 1 year** of dedicated manpower, shared between DESY-IT and the experiments is necessary in order to implement such a project.

Other Projects

An effort has now begun to secure the status of the H1 non-digital documentation. In terms of digital documentation, several initiatives are needed including streamlining the current H1 web content and working with the DESY Library and INSPIRE about future storage of electronic documentation and secondary data. On the more technical side, the database access and reliance of the H1 web server should be examined and a migration of the web server to the DESY-IT virtual environment should be foreseen, relieving the collaboration of future hardware requirements. In order to achieve these goals, and provide H1 with future safe digital documentation,

[5] http://www.teilchenwelt.de
[6] http://quarknet.fnal.gov
[7] http://kworkquark.desy.de
[8] See e.g.: http://www.stanford.edu/group/burchat/cgi-bin/bellis_mediawiki/index.php/Viewpoints_NASA

a total of **1 FTE for 1 year** is required, in addtion to the 1 FTE for 1 year mentioned above in the context of common validation project with DESY-IT. Similar data preservation projects within the ZEUS and HERMES collaborations are still under discussion and FTE estimates are in preparation.

DESY and DPHEP

DESY and the HERA collaborations have played a central role in the international data preservation initiative in high energy physics, DPHEP. This organisation requires a Project Manager to be installed in 2010. Participation in the coordination of various projects at the international level within DPHEP is also possible. Some of the tasks mentioned above can be complemented by a coordination role to be defined within DPHEP during 2011.