The ZEUS data preservation project

Janusz Malka and Katarzyna Wichmann on behalf of the ZEUS collaboration
Deutsches Elektronen-Synchrotron, Notkestr 85, 22607 Hamburg, Germany
E-mail: janusz.malka@desy.de, katarzyna.wichmann@desy.de

Abstract. A project to allow long term access and physics analysis of ZEUS data (ZEUS data preservation) has been established in collaboration with the DESY-IT group. In the ZEUS approach the analysis model is based on the Common Ntuple project, under development since 2006. The real data and all presently available Monte Carlo samples are being preserved in a flat ROOT ntuple format. There is ongoing work to provide the ability to simulate new, additional Monte Carlo samples also in the future. The validation framework of such a scheme using virtualisation techniques is being explored. The goal is to validate the frozen ZEUS software against future changes in hardware and operating system. A cooperation between ZEUS, DESY-IT and the library was established for document digitisation and long-term preservation of collaboration web pages. Part of the ZEUS internal documentation has already been stored within the HEP documentation system INSPIRE. Existing digital documentation, needed to perform physics analysis also in the future, is being centralised and completed.

1. HEP Data Preservation Project
The data collected by High Energy Physics experiments are crucial to our understanding of particle physics. The data preservation effort aims to ensure long-term availability of these data after the end of the experimental collaborations. Data preservation increases the physics potential of experiments, allowing a long-term data analysis, re-using and re-analyzing data, combining results between experiments and using the data for education, training and outreach purposes. Among the many high-energy physics projects the Hadron Electron Ring Accelerator, HERA, holds a special place, since so far it was the world’s only ep collider. The DESY Data Preservation Group (DESY-DPHEP) was established in 2009, shortly after the global DPHEP initiative was launched. For more details about the DPHEP study group please see the homepage http://dphep.org, as well as the first publication of the group “Data Preservation in High Energy Physics”, available at http://arxiv.org/abs/0912.0255 arXiv/0912.255. DESY-DPHEP comprises members from the HERA collaborations (H1, HERMES and ZEUS) in addition to representatives from IT, the DESY Library and INSPIRE. HERA data represent a unique achievement in HEP and thanks to the DESY Data Preservation Project the potential of the data can be further exploited.

2. The ZEUS Data Preservation Project
All of the HERA experiments are now consolidating their respective data analysis models, including finalising the data formats for preservation of the full HERA I+II data. All experiments
Table 1. The DPHEP preservation modes listed in order of increasing complexity.

| Preservation Model | Use Case                                      |
|--------------------|-----------------------------------------------|
| 1. Additional information | Additional information Publication related information |
| 2. Provide data in simplified format | Outreach, training Full scientific analysis possible, based on existing reconstruction |
| 3. Preserve the analysis level software and data format | Retain the full potential of the experimental data |
| 4. Preserve the full simulation and reconstruction software as well as the basic level data |                                         |

are also participating in the development of a long term validation system in cooperation with the DESY-IT division. The current ZEUS analysis model based on MDST (Mini Data Summary Tape) contains a lot of external dependencies and cannot be maintained after 2012 when most of the manpower is going to be reduced. Therefore the ZEUS current analysis software is used to create common usage ntuples (real and MC data). The ROOT-based Common Ntuples content has been iterated over the last few years. These ntuples contain the necessary information to perform all ongoing and planned future analyses, and most of the ongoing analyses are already actively using them. Some low-level information for unplanned new analysis approaches is also included. Simple ROOT ntuple format is used - flat ntuples (no object, no histograms). The resulting total ntuple size is expected to be between 10-20% of the size of MDST data. The storage and access is unchanged with respect to the current model (tapes and dCache). Additionally a stand-alone Monte Carlo, MC, package using existing, frozen executables is also being provided to generate small additional MC sets in the future, foreseeing developments in theory. The different data preservation models established by the DPHEP group, organised in levels of increasing benefits, complexity and cost, are show in table 1. The ZEUS concept fits into the preservation model 3 and 4.

3. Validation framework
In order to preserve the analysis capability for a longer period it would be beneficial to migrate to the latest software versions and technologies for as long as possible, substantially extending the lifetime of the software, and hence the data. It is therefore desirable to develop a framework to automatically test and validate the software and data of an experiment against such changes and upgrades to the environment, as well as changes to the experiment software itself. A generic validation suite, which includes automated software build tools and data validation, provides a solution to this problem that can be shared by different experiments. A prototype version of such a validation framework was successfully installed at DESY-IT during 2010 and now the design, development and implementation of the general version of software preservation framework, called “sp-system”, is now underway. Additionally estimates of the volume of data for preservation from each collaboration are also being prepared, providing the necessary input to the development and evaluation of an archival system to store the HERA data for the long term.

4. Validation of Physics Analyses
A first version of the ZEUS preservation test (validation of the physics analysis of elastic Z0 production) has been integrated into the validation system, successfully testing the data access and ROOT analysis code against different ROOT versions and operating system configurations. Especially event list, cross section, acceptance, invariant mass calculations are checked against
various possible future changes like 32-64 bit machines, new ROOT versions, speed, new operating systems, new data access schemes etc.

5. Stand-alone MC Simulation Package
A preliminary version of the stand-alone MC simulation package specifically designed for this environment has also been prepared. It includes the chain from the generation of the events through simulation of the ZEUS detector and reconstruction up to the Common Ntuple production (see figure 1). This package is a simplified version of the system which is currently used for mass MC production on the GRID. The package contains all necessary dependencies, like calibration, alignment, geometry, frozen executables, steering cards. The other dependencies have been removed. A dedicated interface to new MC generators will also be provided. This scheme is being integrated into the general long term validation framework in collaboration with DESY-IT. In the validation test the log files after each simulation step are checked for warnings and errors. If results are satisfactory the next step is executed and results of checks are stored in test log files for documentation of the test. In the end of test, after the Ntuple production, the content of the Ntuple is validated by comparison with reference Ntuples which were produced with the same detector simulation conditions but for different environment conditions (operating system, ROOT version). A first version has been successfully tested with the virtual SLD5/64-bit reference system and the basic set of distributions were compared using Kolmogorov test in order to make sure that changes of the operating system and/or ROOT version do not change the physics content of the simulation output. The results of this step are stored in the log files as well. In addition a bookkeeping web page for the validation tests has been designed. Such a page contains all necessary information about test results (histograms, log files) and information about operating system and ROOT version. There is still need to develop root based validation tools for MC and software packages on the level of physics analysis.

6. Digital and Non-Digital Documentation
In order to make the preservation of the data and the analysis software useful for future generations, it is necessary to save the vast amount of know-how and information associated with the data and the previous analyses. All HERA experiments have been involved in the reorganisation of non-digital documentation, including relocation to the archive of the DESY Library and digitisation of older documentation. Improvements have been carried out to the content of the web pages of each collaboration and new data management structures have also been investigated. Since the HERA experiments have their beginnings in the mid-1980s, much of this information is kept on non-digital media. Such documentation can provide crucial information in several areas: the background and historic development of the early HERA
physics papers; the evaluation of preliminary results, which were never published; cross-checks, or new evaluation of systematic errors used in previous analyses. Each experiment has a reserved area within the DESY Library Archive where the majority of their non-digital material is now stored. However, in order to make it accessible throughout the collaborations, some of this non-digital information is being brought into digital form, by scanning or photographing.

7. Long-term web server archiving and INSPIRE
The ZEUS collaboration will use long-term web server archiving: a virtual system has now been set up with the DESY WebOffice. Solutions for current web data bases of meetings and public results, both for archiving and future functionality, are being investigated with DESY WebOffice and the INSPIRE/DESY Library teams. There is also continuing progress in the consolidation and revision of ZEUS user-oriented digital documentation, for the preparation of the ZEUS Primer. Finally, several joint projects with INSPIRE and the DESY Library have emerged, concerning storing the internal notes, theses and other documentation.

8. Summary
The most of ZEUS results are already published and most of the ongoing analyses reach the final phase, however the ZEUS experiments is putting strong efforts to give the opportunity to analyse and use these data in the future. The Common Ntuples project will allow to finish currently foreseen analyses, re-analyse them with better tools, and combine them with H1 results. The additional possibility of generation new Monte Carlo samples will make it possible to apply new theoretical predictions.