CEDAR: progress and status report

To cite this article: A Buckley *et al* 2008 *J. Phys.: Conf. Ser.* **119** 052006

Journal of Physics: Conference Series **119** (2008) 052006 doi:10.1088/1742-6596/119/5/052006

CEDAR: progress and status report

A Buckley[1], J M Butterworth[2], L Lönnblad[3], J W Monk[2], E Nurse[2], L Sonnenschein[4], B Waugh[2] and M R Whalley[1]

[1]University of Durham, IPPP, South Road, Durham, England
[2]University College London, Dept. of Physics & Astronomy, Gower Street, London, England
[3]Lund University, Dept. of Theoretical Physics, Sölvegaten 14A, Lund, Sweden
[4]CERN Theory Division, CH-1211 Geneva, Switzerland

E-mail: `cedar@cedar.ac.uk`

Abstract. The CEDAR collaboration has developed and continues to develop a set of software tools for High Energy Physics phenomenology. We outline the status of three of the core CEDAR projects: HepData, a database of experimental measurements; JetWeb, a database and validation tool for Monte Carlo and Rivet, a tool for analysing Monte Carlos in a highly reproducible way.

1. Introduction

Collision events at the Large Hadron Collider (LHC) will take place against a large background of hadronic activity. In order to be able to correctly interpret the physics taking place at the as-yet un-probed high energy scale, it will be necessary to understand the softer underlying event and jet physics.

Monte Carlo event generators are useful tools for describing and understanding both perturbative and non-perturbative features of events. By selecting data from current and past experiments and comparing that data to events generated by different Monte Carlo models and parameter sets, the simulations that most correctly match the anticipated underlying event and jet physics at the LHC can be determined. The χ^2 of the Monte Carlo distributions compared to the data is one measure of the goodness of the fit of the simulation to the data. An example of a scan of χ^2 for Pythia [1] generated events compared to a thrust distribution in the parameter space of Λ_{QCD} and the parton shower mass cut-of M_{min} is shown in figure 1. Figure 1 was produced with 50,000 events per data point using the Professor system for tuning Monte Carlos [2], which interfaces to Rivet (section 4) in order to analyse and generate the Monte Carlo events.

2. HepData

HepData [3] is a core CEDAR service providing data from experiments to other projects such as JetWeb (section 3), as well as a web interface to access the data directly. The original HepData [4] has been maintained as a hierarchical database at Durham for around thirty years. However, this legacy database, which is usually accessed via Fortran routines, lacks many of the features required of a more modern application such as network accessibility, SQL support or a Java interface. For this reason, the CEDAR collaboration has been migrating HepData to a relational MySQL database. The migration is now sufficiently advanced that other CEDAR projects are able to take advantage of the data service offered by the new HepData.

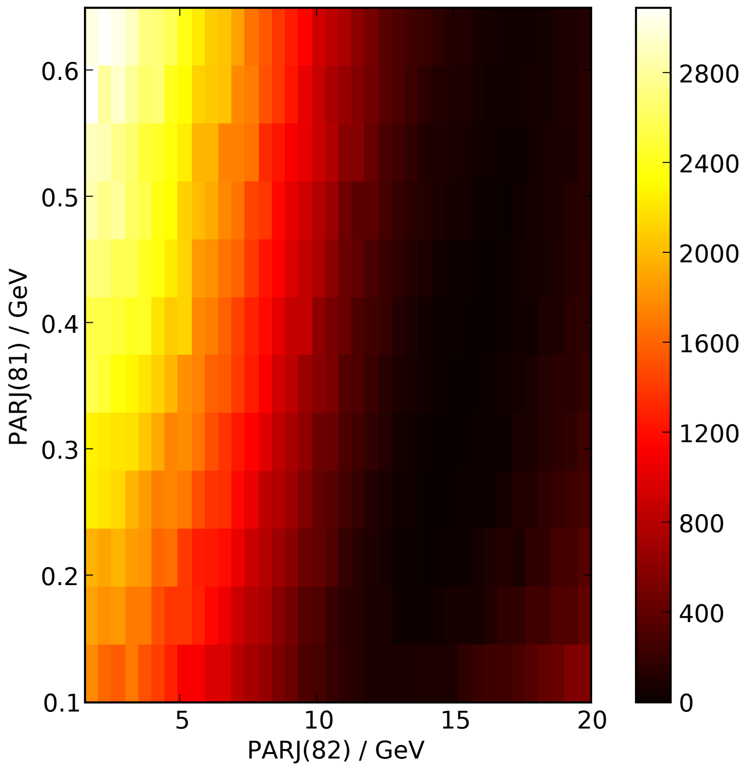

Figure 1. Scan in Λ_{QCD} (y axis) and M_{min} shower cut-off (x axis) parameter spaces of the χ^2 compared to a reference plot. 50,000 events were generated per data point using pythia[1].

A Java object model description of the data contained within HepData now exists and can be used not only by HepData's own web interface that runs inside a Java Tomcat servlet, but can also be incorporated into other Java web applications or programs. The HepData object model uses Hibernate [5] to populate the Java objects with data from the underlying database. Hibernate annotations are used to describe the mapping between objects and data fields and the relationships between different data structures. The only explicit dependence on the underlying database type is in the Hibernate conguration file, so it is trivial to move to a different database, so long as it is supported by Hibernate and contains the correct data structures. The crucial feature of the HepData Java object model is that applications that access HepData solely using the object model are protected from needing to know any of the details of the underlying database.

The data from the object model may also be transported in a variety of other formats. Chief amongst these is the HepML XML format [3]. There are two sub-schemata of HepML in use by Cedar: one is a transportable description of Monte Carlo settings and parameters and the other is a full record of a set of experimental data. It is into the latter of these that HepData can export a full representation of the data hierarchy. HepData also makes data available in AIDA and ROOT formats.

2.1. Database migration

The data held in HepData naturally fits the many to one hierarchical structure of the legacy database well because each data point can only belong to one plot, which in turn can only be held in one dataset, which is contained in one paper. A hierarchical structure has also therefore been developed in the new MySQL database.

In order to migrate the data from the legacy to the new database, the data is first exported from the legacy database as a set of flat text files. The exporting process requires first reading the database using some Fortran code and then organising the records with some Perl scripts.

Journal of Physics: Conference Series **119** (2008) 052006 doi:10.1088/1742-6596/119/5/052006

Each flat file contains all of the entries for every paper for one particular aspect of the data. For example, there are at files that contain all of the Spires ids or all of the x axis values in the legacy database. The flat files are then split up into one file per paper so that the entire database does not need to be scanned in order to retrieve a single paper.

The split flat files can be converted into HepML representations of the data using a Python script, which combines the twenty flat files into a single HepML file. The HepML representation can then be un-marshalled using the Castor object persistency framework [6] to the HepData Java object model. Once in the Java object model, the data is saved to the new database through the Hibernate persistency framework. The data can then be accessed either through the object model, as a HepML file or directly using SQL queries.

3. JetWeb

JetWeb is a web application and database of Monte Carlo results that has been developed by the CEDAR collaboration. JetWeb allows a user to supply a set of Monte Carlo parameters (either input by hand into the web interface or uploaded in a HepML XML file) and perform a search for models that match those parameters. If a matching model already exists within JetWeb's database then a comparison between the Monte Carlo and experimental results can immediately be made across a wide range of datasets from many experiments. The comparison allows a simple visual inspection of the plots or the χ^2 can be calculated between the Monte Carlo and experimental data for all or a subset of the plots. The later is important because it allows one to determine if the Monte Carlo is valid for all interactions, not just those of immediate interest.

If JetWeb's database does not already contain a model that matches the search then results are not returned immediately. Instead, the user is given the opportunity to request generation of Monte Carlo data to match the new model. The data is generated on a computer grid using scripts written out by JetWeb that steer either HZTool [7, 14] or Rivet (section 4).

Where previously JetWeb needed its own internal database of experimental results, it is now able to use the HepData object model to access datasets from HepData to use as comparison data. Other improvements that have been made to JetWeb since the last CHEP conference include an internal generator independent representation of a kinematic cut, which allows JetWeb to break the Monte Carlo jobs into several kinematic regions in order to generate good statistics, and the ability to download comparison and Monte Carlo plots as AIDA histogram files.

3.1. JetWeb output

As an example of how JetWeb may be used, we show two different tunings of the Herwig Monte Carlo compared to photoproduction data from H1. The Jimmy [8] multiple interactions model was used with Herwig [9] and the parameter varied here was `PTJIM`. `PTJIM` is a scale below which there cannot be any emission of radiation during an event, therefore a low value of `PTJIM` allows more radiation than a high value. Values of `PTJIM` of 2 and 3 GeV are shown here, with all other Herwig parameters kept at the default value. Data shown here is taken from two H1 photoproduction papers [10, 11], however a much larger set of data is available from the online JetWeb server [12, 13]. The data are shown in figures 2 and 3. The χ^2 per degree of freedom from the plots shown is given in table 1. Note that Herwig is a leading order Monte Carlo, and as such the shape of the distributions produced is more important than the overall normalisation.

An important difference between [10] and [11] is that in [10] the outgoing lepton is tagged, which allows the fraction of the beam lepton's momentum carried by the photon (y) to be determined. In [11], the lepton is not tagged and y must be estimated using calorimeter cells. This means that while the values of y determined in [10] are not sensitive to `PTJIM`, in [11] a lower value of `PTJIM` ought to favour higher y values due to the increased calorimeter activity. Conversely, the fraction of the photons momentum carried by the parton (x_γ) is dependent

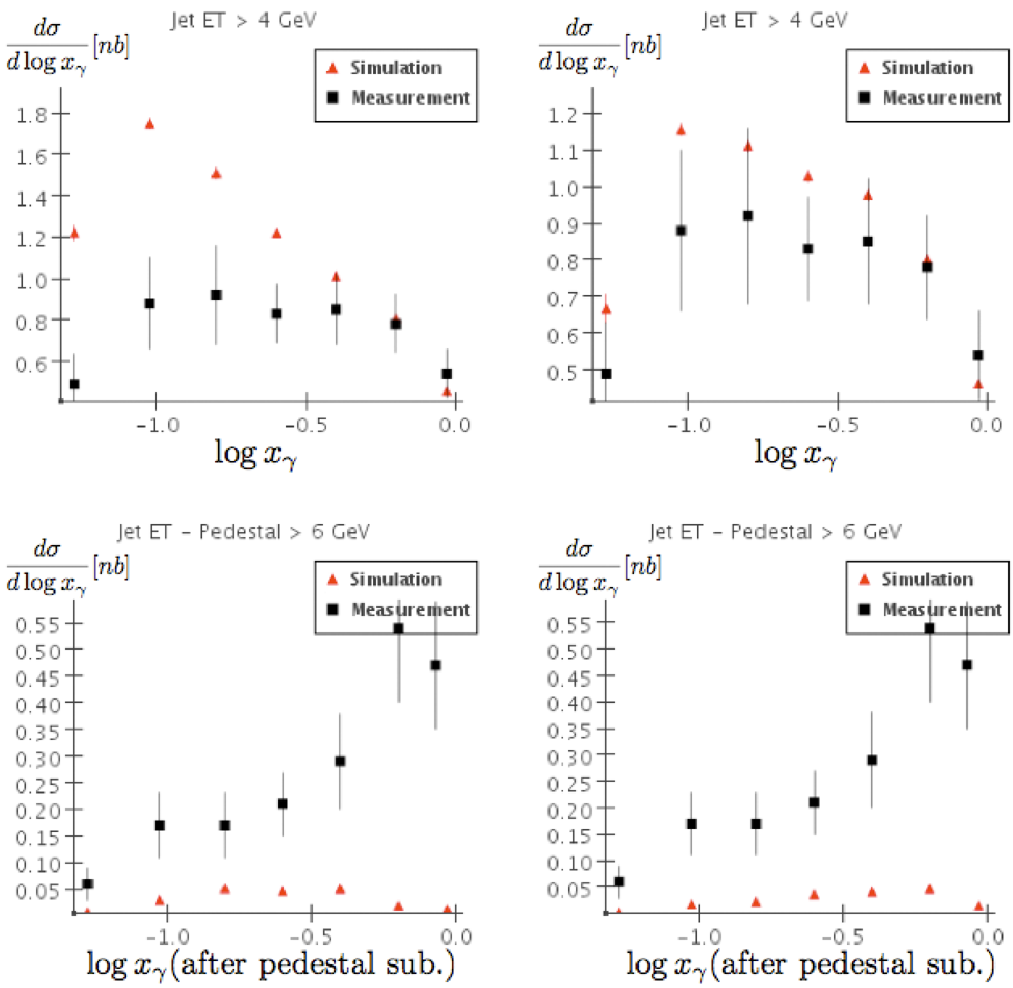

Figure 2. Plots taken from JetWeb comparing Herwig simulation (red triangles) to data (black squares) from [10] for `PTJIM` values of 2 GeV (left column) and 3 GeV (right column). The top plots show x_γ for events with a pair of jets with transverse energy above 4 GeV, the bottom plots show for events with jets above 6 GeV after pedestal subtraction.

Table 1. The χ^2 per degree of freedom from JetWeb models [12, 13] with data shown in gures 2 and 3

	data from [10]	data from [11]	total
`PTJIM`=2 GeV	7.9	6.4	7.4
`PTJIM`=3 GeV	4.67	8.21	6.0

on the jet transverse energies (E_T) and $1/y$ in both papers. The dependence of E_T on `PTJIM` therefore cancels to some extent with the dependence of $1/y$ on `PTJIM` in [11], leading to a reduced `PTJIM` sensitivity of the reported value of x_γ compared to [10].

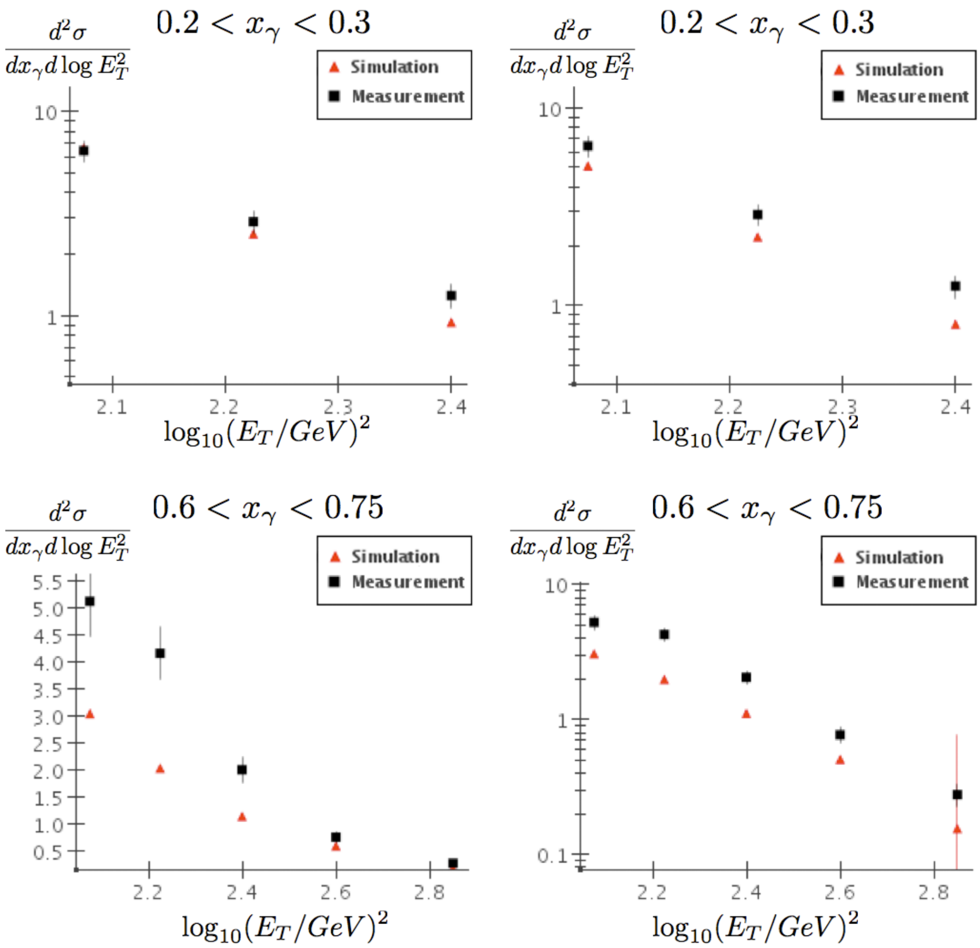

Figure 3. Plots taken from JetWeb comparing Herwig simulation (red triangles) to data (black squares) from [11] for `PTJIM` values of 2 GeV (left column) and 3 GeV (right column).

The shape of the Monte Carlo distributions produced for [11] is a closer match to the shape of the data compared to the distributions for [10]. A clue as to why this might be the case can be gained from the lower row of plots in figure 2. In these plots the average out of cone radiation per unit of $\Delta\eta\Delta\phi$ multiplied by the cone area (called the jet pedestal) has been subtracted from the jet E_T , which is then required to be above 6 GeV. Comparing the lower plot in figure 2 to the upper plot, in which there is no pedestal subtraction and the minimum jet E_T is 4 GeV, indicates that at a jet E_T of 4 GeV the jets simulated by Herwig and Jimmy are dominated by radiation from the underlying event. The distributions in [11], on the other hand, have a minimum jet E_T of 7 GeV, and as such will not be affected as much if there is too much soft radiation from the underlying event.

4. Rivet

Rivet [7], the Robust Independent Validation of Experiment and Theory, is being developed as a C++ replacement for the HZTool [7, 14] Fortran package and is currently at release version 1.0b1, available from the Rivet project page [15]. HZTool was originally developed by the H1 and

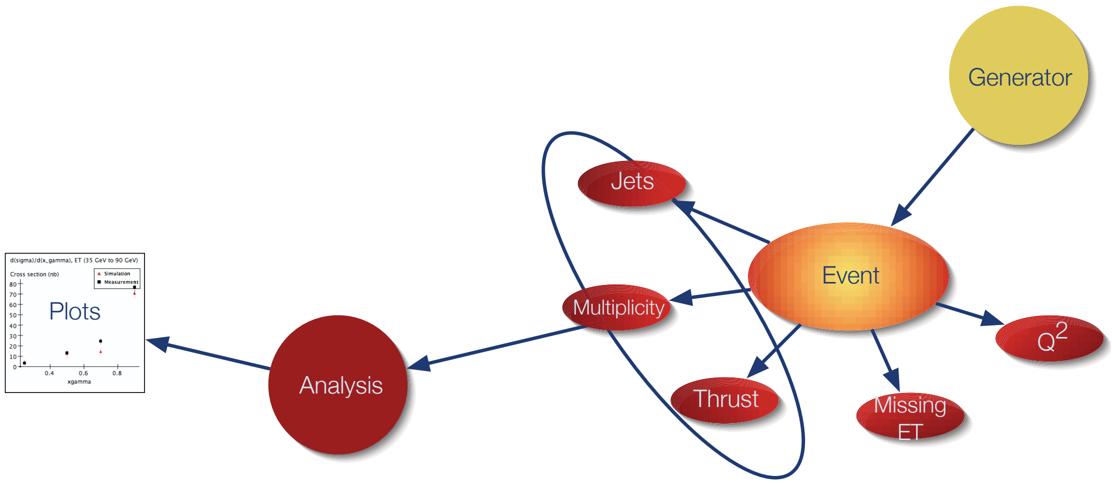

Figure 4. Schematic of a Rivet and RivetGun analysis chain. Events are generated with a generator and passed to generator independent projections. The projections project useful quantities, which are then grouped together in analyses. The analyses output histograms.

Zeus collaborations to allow them to reproduce each others analyses on Monte Carlo simulated events. Because HZTool contains subroutines and functions common to many analyses, it is now used to reproduce analyses from other experiments such as those at LEP or the Tevatron as well as the original HERA experiments.

The key innovation within Rivet over HZTool is the concept of a projection. A projection is an object that can take a HepMC Monte Carlo event record and project from it a quantity or set of quantities that are used in an analysis. For example, the simplest projection is the `FinalState` projection, which makes available the set of four-vectors representing only those stable final state particles that are long lived and reach the detector. The `FinalState` projection is therefore used by most other projections. A more complicated projection is the `D0ILConeJets` projection, which makes available the jets found by the cone-jet algorithm used by the DØ experiment during the Tevatron Run II [16]. An analysis simply consists of a set of projections together with event selection criteria (figure 4).

A list of projections to be run and the order in which to run them is constructed by Rivet at runtime and each projection is projected only once per event, even if it is used by many other projections and in many separate analyses. This greatly reduces the execution time if multiple analyses are run simultaneously. The current (as of September 2007) release of Rivet provides four analyses that match measurements from the Zeus [17], DØ [18], H1 [19] and Aleph [20] experiments. Other analyses can be implemented by the user without modifying the Rivet library.

Rivet can produce histogram output in both the AIDA and Root file formats and can also read data from an AIDA file. Histogram booking routines are included that can optionally reproduce the binning used by another AIDA histogram that is read from a file. A set of these AIDA files containing comparison experimental data from HepData (section 2) for the supplied Rivet analyses are included in the Rivet release tarball.

4.1. RivetGun

When compiled with the included automake scripts, Rivet produces a set of dynamically linked libraries that contain all of the generator independent projections and analyses. In order to run an analysis, however, these routines must be interfaced to a Monte Carlo event generator. RivetGun is a steering package for Rivet that interfaces the Rivet analyses to many Monte Carlo generators and allows them to be steered through a simple command line interface. The current list of supported generators is: both the Fortran and C++ implementations of Herwig; the Fortran and C++ implementations of Pythia; Alpgen; Charybdis and Sherpa. The full set of generators available on any given system can be found by running `rivetgun -h`. RivetGun can pass generator-specific parameters to a generator through the use of the command line interface with the option `-p parm=value`. The generator and analyses to be run, as well as the beam particles, momenta and number of events to be generated, can also be steered from the command line. It is also possible to run RivetGun without running Rivet by using the -R option, in which case RivetGun becomes a very nice way to steer a Monte Carlo generator even without the Rivet functionality.

5. summary

The tools under development by the CEDAR collaboration are now suffciently advanced to be useable by other projects and groups both inside and outside of CEDAR. The new HepData object model and database structure is complete and the migration process from the old to the new database is well established. The significant remaining tasks are to finalise development of the Tomcat web application front end and actually migrate all of the available data.

JetWeb has been internally restructured to use the new HepData and is currently doing so where the relevant data exist in the new database. Example JetWeb output was shown, which demonstrates the ease with which Monte Carlo results can be inspected. JetWeb is available for use online, with the caveat that it is under development.

A beta version of the Rivet Monte Carlo analysis framework has been released bundled with four analyses and appropriate comparison datasets. The central Rivet concept of a projection is very elegant and makes it as simple as possible for a user to write a new analysis. Indeed, the main feature lacking from Rivet is a larger number of useful analyses. The reader is therefore encouraged to use Rivet and either contribute their own analysis or provide a wish-list of analyses and projections as feedback.

Acknowledgments

The CEDAR collaboration thanks the UK Science & Technology Facilities Council (STFC) for their generous support.

References

[1] T. Sjöstrand, S. Mrenna and P. Skands, JHEP **0605** (2006) 026 [arXiv:hep-ph/0603175].
[2] The Professor tuning system:
`http://projects.hepforge.org/professor`
[3] A. Buckley, W. J. Stirling, M. R. Whalley, J. M. Butterworth, J. Monk, E. Nurse and B. Waugh, arXiv:hep-ph/0605048.
[4] HepData reaction database:
`http://durpdg.dur.ac.uk/hepdata/reac.html`
[5] Hibernate:
`http://www.hibernate.org/`
[6] Castor framework:
`http://www.castor.org`
[7] B. M. Waugh, H. Jung, A. Buckley, L. Lönnblad, J. M. Butterworth and E. Nurse, arXiv:hep-ph/0605034.
[8] J. M. Butterworth, J. R. Forshaw and M. H. Seymour, Z. Phys. C **72**, 637 (1996) [arXiv:hep-ph/9601371].
[9] G. Corcella *et al.*, JHEP **0101**, 010 (2001) [arXiv:hep-ph/0011363]; hep-ph/0210213.
[10] C. Adloff *et al.* [H1 Collaboration], Phys. Lett. B **483**, 36 (2000) [arXiv:hep-ex/0003011].
[11] C. Adloff *et al.* [H1 Collaboration], Eur. Phys. J. C **1**, 97 (1998) [arXiv:hep-ex/9709004].

Journal of Physics: Conference Series **119** (2008) 052006 doi:10.1088/1742-6596/119/5/052006

[12] JetWeb model 8:
`http://jetweb.cedar.ac.uk/jetweb-webapp/JWSearch/$mdl=8&`

[13] JetWeb model 9:
`http://jetweb.cedar.ac.uk/jetweb-webapp/JWSearch/$mdl=9&`

[14] J. Bromley *et al.*, Prepared for Workshop on Future Physics at HERA (Preceded by meetings 25-26 Sep. 1995 and 7-9 Feb. 1996 at DESY), Hamburg, Germany, 30-31 May 1996

[15] Rivet project:
`http://projects.hepforge.org/rivet/`

[16] G. C. Blazey *et al.*, arXiv:hep-ex/0005012.

[17] S. Chekanov *et al.* [ZEUS Collaboration], Eur. Phys. J. C **23**, 615 (2002) [arXiv:hep-ex/0112029].

[18] V. M. Abazov it et al. [D0 Collaboration], Phys. Rev. Lett. **94**, 221801 (2005) [arXiv:hep-ex/0409040].

[19] S. Aid *et al.* [H1 Collaboration], Phys. Lett. B **356**, 118 (1995) [arXiv:hep-ex/9506012].

[20] D. Decamp *et al.* [ALEPH Collaboration], Phys. Lett. B **273** (1991) 181.