CEDAR

Data Storage and Model Validation in Particle Physics

Jonathan Butterworth

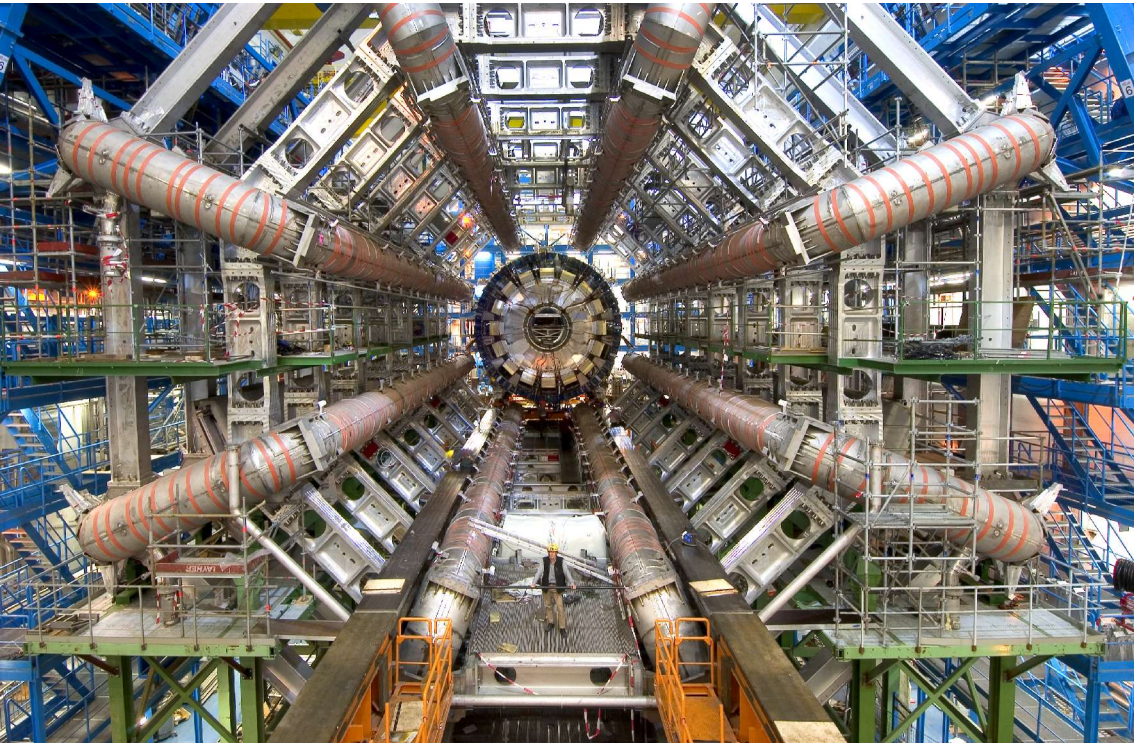
UCL High Energy Physics

DCC Information Day

16th Jan 2007 UCL

CEDAR

Particle Physics Data



- Most current particle physics data comes from (very) large experiments.
- e.g. ATLAS > 3000 Physicists.
- Raw data (after an online preselection) comes in at a rate of ~1 PetaByte per year.
- Also need similar sizes samples of simulated and reconstructed data.

CEDAR

Particle Physics Data

- Experiments usually have a large suite of their own software, which is required in order to read & reconstruct their raw data
  - maintenance nightmare!
- Experiments usually have a lifetime of O(10) years
  - encourages some robustness and maintainablility
  - at least >1 person understands the system!
- To reconstruct the data, many calibration constants and other expert information are required
  - Usually an effort to have a “final reconstruction” and store the reconstructed data. Big challenge!
    - Is enough information stored to make it useful
    - Can it still be read?
    - Last one out, turn out the lights...

CEDAR

Particle Physics Data

- There have been examples of good practice
  - e.g. Publications still from reanalysis 1980's data from DESY
- LEP data (<2000) still being analysed.
  - Will this continue with new people?
- What will happen to Tevatron and HERA data (finishing 2009, 2007)?
  - To be decided.

CEDAR

Related Problem

- Publications often involve comparison of complex variables to the results of detailed simulations.
- Can the simulations be redone if a new model comes along?
  - Often difficult.
  - Sometimes not enough information in the paper.
- Result is an under use of old data.

CEDAR

Solutions

- CEDAR (Combined Escience Data Analysis Resource)
  - PPARC (and recently EU) funded set of software tools, the main aim of which is to couple validation tools for *physics Monte Carlo programs* (and other physics calculational) tools with *data.*
  - Durham + UCL
  - Code (Rivet) for repeating analysis on MC
  - Database (HepData) archive of high energy physics data
  - Portal and Library (JetWeb) for generating, storing and searching comparisons
- Also provides (since it needs them itself)
  - Lightweight code development environment (HepForge).
  - XML descriptions of HepData records and generator parameters (HepML, in collaboration with LCG Generators)

CEDAR

HepForge

- Handy, lightweight development platform for small reusable HEP software projects.
- Important for archiving of maintained & validated software
- Pick and mix of services available:
  - Version control, bug/issue tracking, download/release management, web space, wiki, mailing lists, shell account...
  - Used by about 25 smallish software projects
- More info at
  - http://hepforge.cedar.ac.uk (or http://hepforge.org)

CEDAR

HepData

- Data store for all HEP measurements
  - Has been in existence >20 years
  - Data from >1400 experiments
  - Correlated and linked online to the SPIRES publication database (SLAC)
- Current public version is in legacy Berkeley DB
- CEDAR have migrated it to a MySQL DB
  - long term maintenance
  - java object model
  - accessible to jetweb and others

CEDAR

Rivet

- Robust Independent Validation of Experiment & Theory
- Approximately equivalent to a C++ replacement of pre-existing Fortran (HZTool, >10 years old, quite widely used)
  - Will make use of some existing external libraries (CLHEP, KtJet/FastJet etc)
- Rivet is independent of the physics simulation being used
  - performs analysis on a set of particles from a simulated collision.
- Outputs histograms for comparison to data from HepData and for inclusion in JetWeb.

CEDAR

RivetGun

- RivetGun interfaces Rivet to a variety of physics programs which model high energy collisions.
- Plan to allow configuration of generators using HepML.
- Intended as a convenient standalone tool for MC developers, experimentalists and theorists, as well as a component of JetWeb.

Rivet and RivetGun

- Design and development ongoing
  - Framework there, but not much physics content yet.
  - First beta release with some useful functionality expected in the next few weeks.
- See http://hepforge.cedar.ac.uk/rivet and http://hepforge.cedar.ac.uk/rivetgun

CEDAR

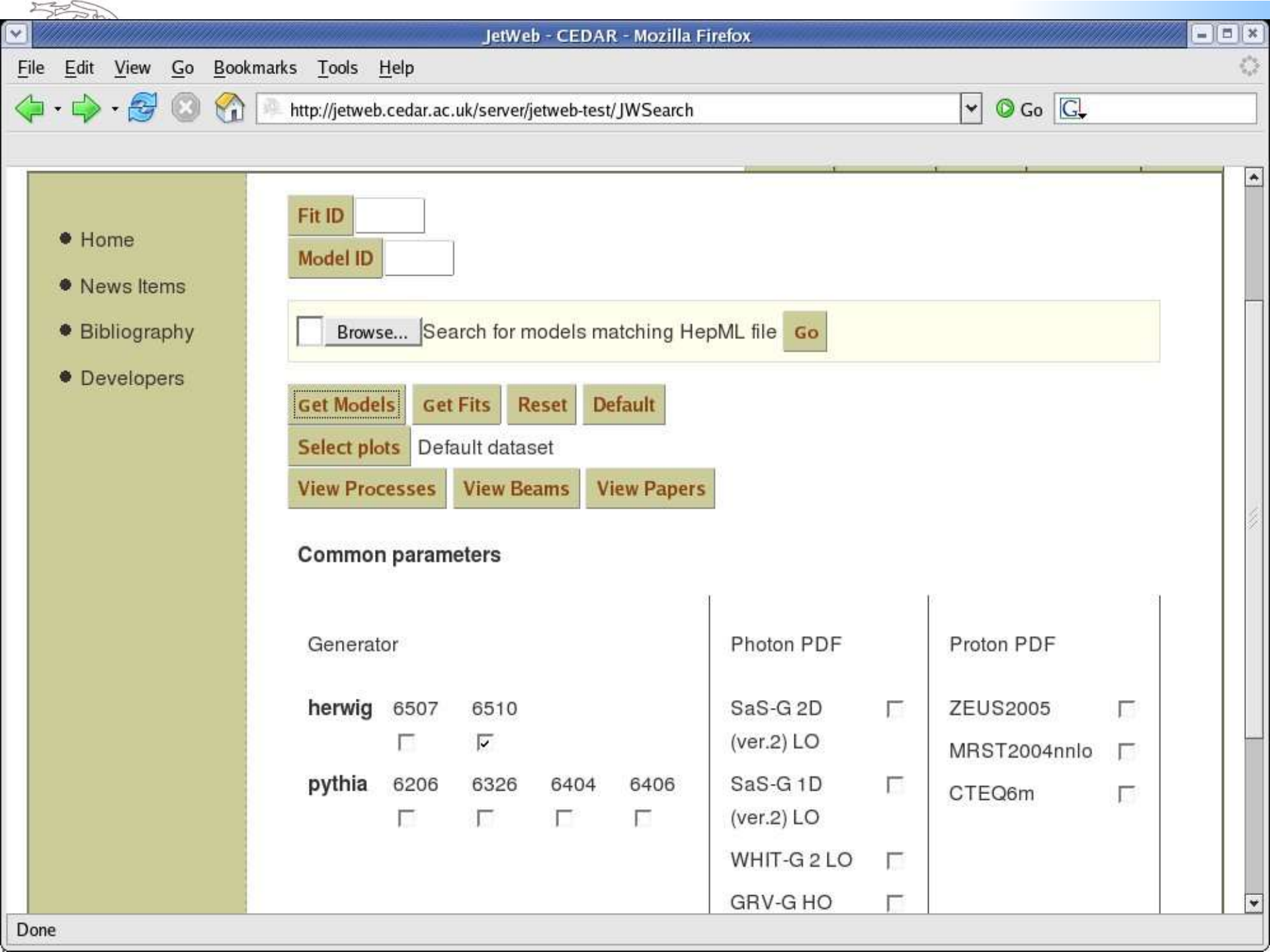
JetWeb

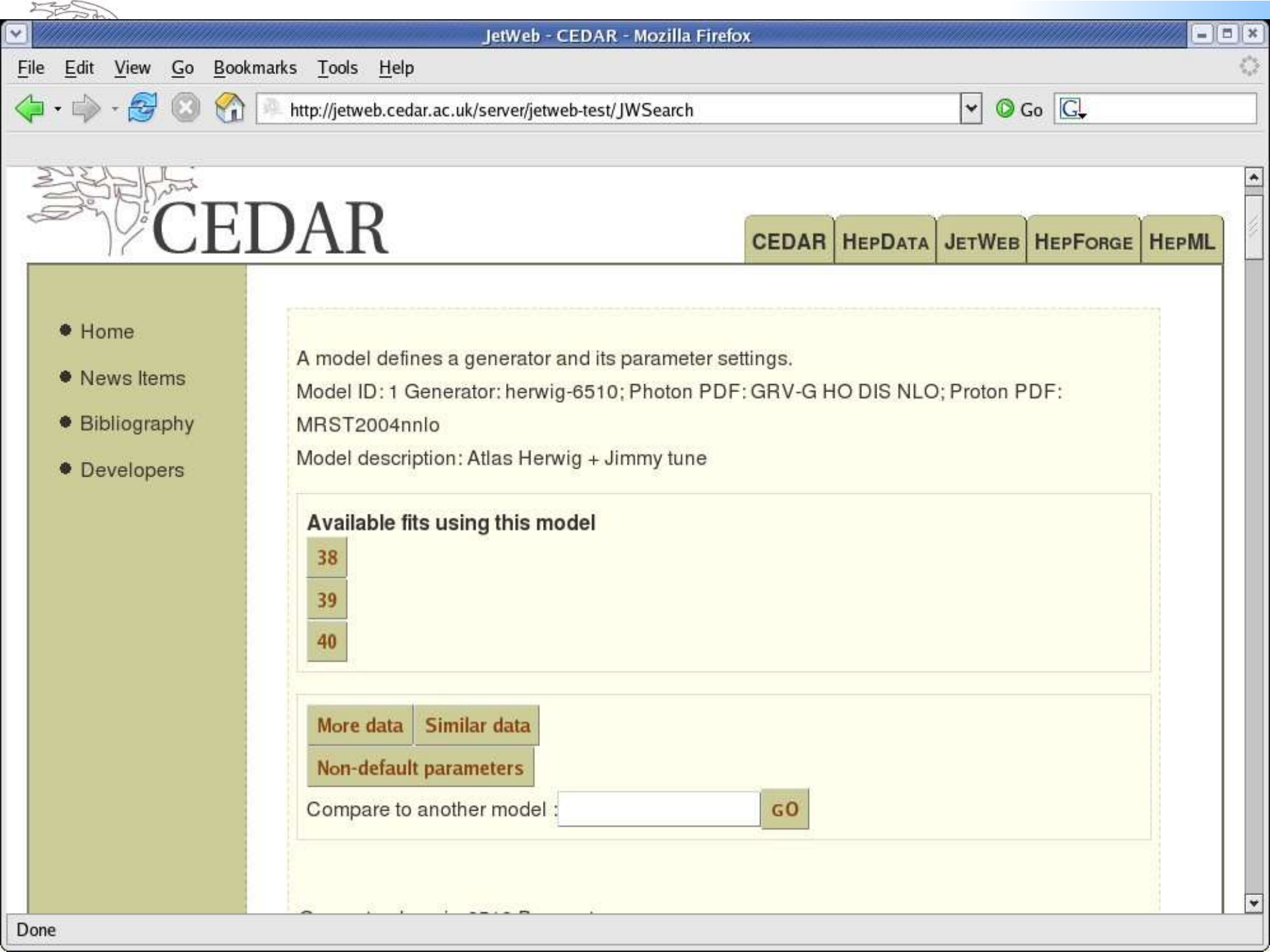
- Web and database server for archiving validated MC models.
- Use Rivet and its precursers, running on LCG, to generate simulated data analyses.
- Use HepML for describing validated physics models
- Use HepData as the single source for all measurements

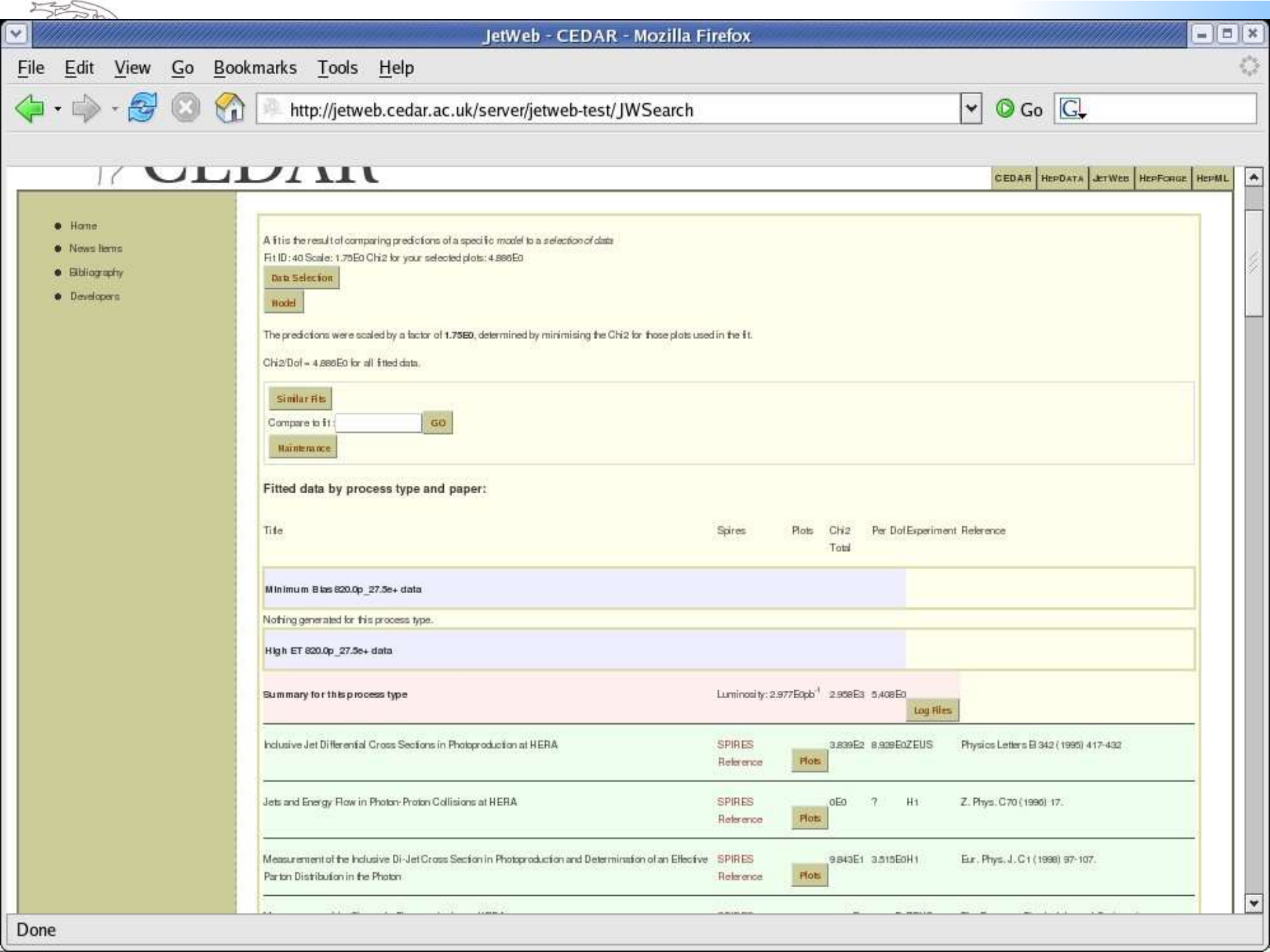
CEDAR

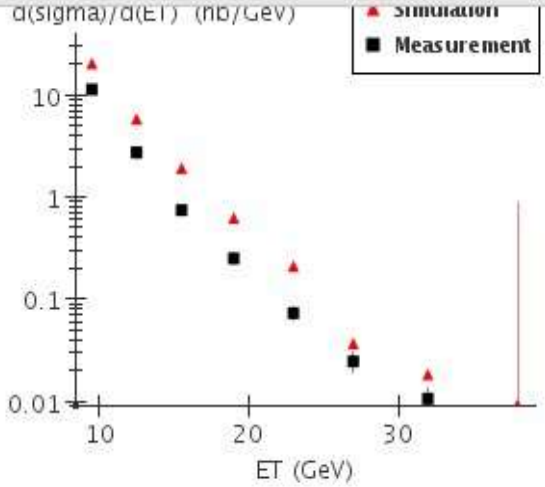
JetWeb Goals

- Build up database of validated models using wide range of existing data
- Add new *physics generators* and *data* rapidly as they appear.
- Users add their own new parameter settings.
- Add more user front-end facilities for interactive tuning and analysis.
- See http://jetweb.cedar.ac.uk


JetWeb - CEDAR - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://jetweb.cedar.ac.uk/server/jetweb-test/JWSearch
Go
Home
News Items
Bibliography
Developers
Fit ID
Model ID
Browse...
Search for models matching HepML file
Go
Get Models
Get Fits
Reset
Default
Select plots
Default dataset
View Processes
View Beams
View Papers
Common parameters
Generator
herwig 6507 6510
pythia 6206 6326 6404 6406
Photon PDF
SaS-G 2D (ver.2) LO
SaS-G 1D (ver.2) LO
WHIT-G 2 LO
GRV-G HO
Proton PDF
ZEUS2005
MRST2004nnlo
CTEQ6m
Done


JetWeb - CEDAR - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://jetweb.cedar.ac.uk/server/jetweb-test/JWSearch
Go
CEDAR
CEDAR HEPDATA JETWEB HEPFORGE HEPML
Home
News Items
Bibliography
Developers
A model defines a generator and its parameter settings.
Model ID: 1 Generator: herwig-6510; Photon PDF: GRV-G HO DIS NLO; Proton PDF: MRST2004nnlo
Model description: Atlas Herwig + Jimmy tune
Available fits using this model
38
39
40
More data
Similar data
Non-default parameters
Compare to another model :
GO
Done


JetWeb - CEDAR - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://jetweb.cedar.ac.uk/server/jetweb-test/JWSearch
Go
CEDAR
CEDAR HepData JetWeb HepForge HepML
Home
News Items
Bibliography
Developers
A fit is the result of comparing predictions of a specific model to a selection of data
Fit ID: 40 Scale: 1.75E0 Chi2 for your selected plots: 4.886E0
Data Selection
Model
The predictions were scaled by a factor of 1.75E0, determined by minimising the Chi2 for those plots used in the fit.
Chi2/Dof = 4.886E0 for all fitted data.
Similar Fits
Compare to fit:
GO
Maintenance
Fitted data by process type and paper:
Title
Spires
Plots
Chi2 Total
Per Dof
Experiment
Reference
Minimum Bias 820.0p_27.5e+ data
Nothing generated for this process type.
High ET 820.0p_27.5e+ data
Summary for this process type
Luminosity: 2.977E0pb-1
2.958E3
5.408E0
Log Files
Inclusive Jet Differential Cross Sections in Photoproduction at HERA
SPIRES Reference
Plots
3.839E2
8.929E0
ZEUS
Physics Letters B 342 (1995) 417-432
Jets and Energy Flow in Photon-Proton Collisions at HERA
SPIRES Reference
Plots
0E0
?
H1
Z. Phys. C70 (1996) 17.
Measurement of the Inclusive Di-Jet Cross Section in Photoproduction and Determination of an Effective Parton Distribution in the Photon
SPIRES Reference
Plots
9.843E1
3.515E0
H1
Eur. Phys. J. C1 (1998) 97-107.
Done



Chi2 Contribution: (chi2 / DoF): 5.236E1 / 6E0

Data (black) was scaled by: 1.2E0
The model was scaled by 1.75E0

Vector output of plotted data

For this fit, the simulated data was generated as

ID: 2 High ET in
820.0 GeV p - 27.5 GeV e+ collisions. More

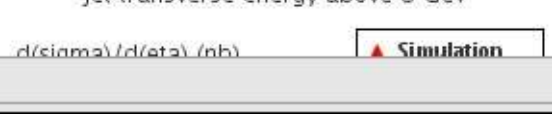
More

Pull for each point:

(3.25E-1) (2.279E0) (2.677E1) (9.924E0) (1.306E1) (0E0) (0E0) (1E-4)

(this plot not included in the fit)

Jet transverse energy above 8 GeV



CEDAR

HepML

- XML Schemas to allow
  - exchange of parameters needed to reliably reproduce a particular generator run.
  - exchange of HepData records.
- Collaboration with LCG Generators (MCDB subproject).

CEDAR

Summary

- Particle physics experiments produce a lot of data
- Some of it needs to be kept available for re-analysis
  - Maintenance of complex experimental software probably not a long term solution
  - Producing a “last word” reduced data set (final calibrations, enough flexibility to measure new things) is probably the best way forward
  - Important sociolgical and technical issues still to be addressed.
- Some of it is published in the form of distributions
  - Detailed knowledge of how they were made must be preserved
  - Storing this, and the data, online means more use is made of it.