Preservation of HERA data and options for HERA data re-analyses

Andrii Verbytskyi

DESY

February 18, 2019

HERA data

HERA reminder:

- The only $e^{\pm}p$ collider, 1991-2007;
- $27.5 GeV\ e^{\pm}$; $460, 575, 820, 920 GeV\ p$;
- (Un)polarized $e^{\pm}$ collide with p;
- Polarised $e^{\pm}$ collide with $H/D/.../Xe$ targets;
- p collide with nuclear targets.

Motivation for preservation:

- Future data (re-)analysis with new models and new approaches.
- Modeling for the future experiments.

What is preservation

- Data Preservation is **NOT** about bytes, programs and files, even if these components are important.
- Data Preservation is about abilities to produce **physics results**. Requires knowledge, eagerness and **manpower**.

Use cases for physics with HERA data (DIS-2016 raw copy)

- Something that now we are not aware about.
- QCD:
  - Proton structure, e.g. F_2 and F_L, strangeness in the proton;
  - Diffraction, e.g. combination of measurements;
  - Jets and event shapes with NNLO;
  - Photon structure, instantons, pentaquarks, etc.
- EW physics:
  - Prompt photons;
  - Electroweak couplings.

See arXiv:1601.01499 and arXiv:1512.03624 for details.

Selected H1 and ZEUS papers and preliminaries since 2016

- **Something that now we are not aware about.**
  ZEUS: Azimuthal particle correlations as a probe of collectivity in deep inelastic electron-proton collisions at HERA, zeus-prel-18-001
- **Strangeness in the proton**
  ZEUS: Strange content of the proton from charm in CC, ICHEP2018 talk
- **Diffraction**
  ZEUS: Studies of the diffractive photoproduction of isolated photons at HERA Phys. Rev. D 96 (2017) 032006
- **Jets with NNLO**
  H1: Determination of the strong coupling constant alphas(MZ) in next-to-next-to-leading order QCD using H1 jet cross section measurements, Eur.Phys.J.C77 (2017), 791
- **Pentaquarks**
  ZEUS: Search for a narrow baryonic state decaying to pKS0 and pbar KS0 in deep inelastic scattering at HERA, Phys. Lett. B 759 (2016) 446
- **Prompt photons**
  ZEUS: Further studies of isolated photon production with a jet in deep inelastic scattering at HERA, JHEP 1801 (2018) 032
- **Electroweak couplings**
  H1: Determination of electroweak parameters in polarised deep-inelastic scattering at HERA, Eur.Phys.J.C78 (2018), 777

Very accurate predictions on the topics. The remaining topics will be covered in the next couple years.

Technical part

Data: H1+ZEUS

All H1 and ZEUS data are available in DESY and in MPP.
MPP bits preservation is similar to approach from DESY.
The main differences comes from the ideas to

- Enable option for worldwide access via Grid.
- Study options to benefit from larger Data Preservation efforts.

Software: ZEUS

- Main software for the analysis is vanilla ROOT.
- "Analysis" does not include reconstruction, i.e. reconstruction software is frozen and is used for MC production only.
- Additional software includes:
  - ZEVIS, the event display based on ROOT;
  - CNINFO, the event data base, based on ROOT and SQLite3;
  - ZMCSP Monte-Carlo standalone generation and reconstruction packages.
- +any ROOT extension that will work for you...

Software environment:ZEUS

A certain environment is needed for the analysis. As of 2019 the demands are low and easy to fulfill:

- DESY provides an access to a batch computing cluster.

In parallel:

- Virtual machines(VM) looks like a very attractive **long-term** solution;
- The way other experiments (LEP/LHC) are going.

Because of very generic requirements it is foreseen that both environments will remain functional for a long time.

Software: H1

- Main software for the analysis is ROOT with custom classes.
- Full chain of software can be recompiled. The analysis chain can start from raw data.
- Recently became available:
    - Event display

Documentation: H1+ZEUS

Coverage:

- Documentation on the data;
- Experiment policy on data access and usage;
- Manual for a possible analysis;
- Manual for the MC generation, including new MC generators (ZEUS only);
- Statements on dedicated resources;
- . . .

From the point of view of physics: **The documentation with enough information for an estimation of particular analysis opportunity with the preserved data. Note the available active experts in house: I. Abt, D. Britzger, A. Caldwell, V. Chekalian, A. Vebytskyi + H. Abramowicz**

ZEUS+H1 Data Preservation summary

(Un)polarized $e^{\pm}$ are collided with protons.

- Data are accessible in DESY and MPCDF.
- Documentation is stored in DESY library/Inspire/web-server;
- Analysis requires only standard software (ROOT) with little (H1) or no (ZEUS) custom classes;
- An option for MC production with new and old MC generators exists;
- Virtualization (ZEUS only)+ dedicated DESY machines+ access to batch cluster.
- Grid VO resources.
- Manpower is the key issue: New collaborators are welcome to analysis!

What that has to do with MCEGs and related software

- Many phenomenology studies either with raw data or published results rely on MC, which requires some manpower for production.
- Older collaborations have shortage of manpower.
- More manpower required for software → less manpower for physics.
- **Properly working software is important**

Rivet as software (and examples of other codes for context)

The definition of properly working software

Define by example: Take two examples of most popular SW packages in physics and outside of Physics.

- Pythia. Pythia6 – 7 years after last release, still in usage, documentation and codes are available, can be compiled with most FORTRAN compilers on most systems. Pythia8 – completely standalone generator, not relying on any external library. Can be compiled with most c++98 compilers on most modern systems. Excellent documentation. Both packages were in standard RedHat and Debian repositories.
- sqlite3 Second most installed software in the world. Quote " Hipp, Wyrick & Company, Inc., (Hwaci) is a small company but it is also closely held and debt-free and has low fixed costs, w which means that it is largely immune to buy-outs, take-overs, and market down-turns. Hwaci intends to continue operating in its current form, and at roughly its current size until at least the year **2050**. We expect to be here when you need us, even if that need is many years in the future. "

User experience: Installation

Code	Time	Lang	Build	Run	Opt	Binaries from reps. EL7/EL6/Mac /Win/Ubt	Dependencies from reps. EL7/EL6/Mac /Win/Ubt	Relocatable	Install. bugs
sqlite3	1	1/0	c autotools			yes/yes/yes/yes/yes	yes/yes/yes/yes/yes	yes	No
Pythia8	1	1/0	c++98		5	yes/yes/yes/no/yes	yes/yes/yes/yes/yes	yes	No
fastjet/fjcore	10	1/0	c++0x?		10	no/yes/yes/no/yes	yes/yes/yes/yes/yes	yes	No
root6	5	1/0	c++0x		100	yes/yes/yes/yes/yes	yes/yes/yes/yes/yes	no	No
SHERPA-MC	10	1/0	c++0x autotools		10	no/no/no/no/no	yes/yes/yes/no/yes	yes	No
Herwig7	60	1/0	c++14 boost ThePEG autotools lhapdf	ThePEG boost, lhapdf fastjet	10	no/no/no/no/no	no/no/no/no/no	no	Yes
lhapdf6	10	2/1	c++11 Cython autotools python2	python2	0	yes/no/no/no/no	yes/no/yes/no/yes	yes	Yes
Rivet	20	4/2	c++11 Cython python2 YODA autotools fastjet	python2 latex fastjet	0	no/no/no/no/no	no/no/yes/no/yes	yes	Yes

- Time – install time in minutes for minimal functional package using best option.
- Lang. – number of required languages to build and run SW.
- Build – packages needed to build.
- Opt. – number of optional subpackages.

User experience: Reliability and support

Code	Dev.	Known	Support	Response,	Future data	Average format	API
	Team	Bugs		days	format compat.	lifespan	lifespan
					input/output	input/output	
sqlite3	Hwaci	No	ML,Tel$,P$	$<$ 1	50 y/50 y	20 y/20 y	10 y
Pythia8	10	No	ML	1-10	Unk/Unk	15 y/30 y	5-10 y
fastjet/fjcore	5	No	ML,P	1	N/A	N/A	5-10 y
root6	CERN	Yes	Forum	1	25 y/25 y	25 y/25 y	10-15 y
SHERPA-MC	20	Yes	BT	1	Unk/Unk	10 y/30 y	Unk
Herwig7	20	No	ML	1-10	Unk/Unk	5 y/30 y	Unk
lhapdf6	5	Yes	ML	10-100	Unk	10 y	1-5 y
Rivet	5	Yes	ML	1-1000	Unk/Unk	2-3 y	0.1-1 y

- P – personal
- $ – paid
- ML – mail lists
- BT – bg tracker
- Note: for some data in table the statistics might be not high enough

Popularity and usage

Code	Usage area	Target group	Users Active	Users once+	Market share	Competitors	Competitors state	Adoption
sqlite3	HEP, Science, Industry	Humanity	$>10^9$	$>10^9$	Dominant	>100	OpenSource, Alive	Vol.
Pythia8	HEP	HEP pheno, analysis, simulation	500	10000	10%+	10	OpenSource, Alive	Vol.
fastjet/fjcore	HEP	HEP pheno, analysis	500	5000	Dominant	5	All dead	Vol.
root6	HEP, Science, Industry	HEP analysis	$>10^4$	$>10^5$	50%+	10-100	All dead in HEP	Vol./Forced
SHERPA-MC	HEP	HEP pheno, analysis, simulation	50	200	10%+	10	OpenSource,Alive	Vol.
Herwig7	HEP	HEP pheno, analysis, simulation	50	200	10%+	10	OpenSource,Alive	Vol.
lhapdf6	HEP	HEP pheno, analysis, simulation	500	5000	50%+	10	Private versions, forks	Vol.
Rivet	HEP	HEP simulation	50	500	50%+	2	OpenSource,Alive	Vol./Forced

- Vol. – volunteer adoption by users
- Note: The numbers for HEP SW are (educated) guesses only, given in order of magnitude. You are free to disagree with them.

Rivet as software: Overview

Very often physics SW require enormous amount of efforts and expertise to use them. **A lot of efforts are reducible**.
In some cases the costs/benefits ratio of using software is unacceptably high or even negative.
For smaller groups and/or older experiments: the high portability, easy and standard install, high compatibility of input and output formats, are needed to spend less time on software and more on physics.

Rivet should be improved significantly before it could be used with positive net benefit by older experiments. Most issues depend fully on Rivet authors and cannot be solved by anyone outside. Moreover, experiments have no manpower to do it (see Stefans talk).

Rivet vs. data preservation

Rivet vs. data preservation

Contra:

- Rivet as a "data preservation" or "analysis preservation" is an obfuscation of the idea of data preservation.
  - "If one can use Rivet, then ZEUS and H1 preserved data can be trashed." – be sure, there will be ideas like that
  - "If one can use Rivet, no expertise, no efforts, workforce and funding etc is needed for old experiments." – natural consequence
- "Preservation" of analysis implies existence of the subject in past. Rivet was not used in ZEUS or H1. **If to go this way, HZTOOL or HZTOOLRivet are prefferable!**
- Introduction of official "analyses" **by ZEUS/H1** would require a lot of manpower that could be used to do physics. Moreover, even if the official analyses would be introduced, there are high chances that
  - Changes in Rivet API will make them unusable in 1-2 months
  - "Improvements" in Rivet are extremely frequent and the experiments will not be able to guarantee correctness of "improved" codes

Pro:

- Wider usage of HERA data in phenomenology studies.

Rivet 2.7.0

```
In file included from ../../include/Rivet/Tools/BinnedHistogram.hh:6:0,
                 from ../../include/Rivet/Analysis.hh:15,
                 from Analysis.cc:3:
../../include/Rivet/Tools/RivetYODA.hh: In instantiation of âĂŸbool Rivet::bookingCompatible(
     TPtr, TPtr) [with TPtr = std::shared_ptr<YODA::Histo1D>]âĂŹ:
../../include/Rivet/Analysis.hh:1001:56:   required from âĂŸstd::shared_ptr<_Tp1> Rivet::
     Analysis::addOrGetCompatAO(std::shared_ptr<_Tp1>) [with AO = YODA::Histo1D]âĂŹ
Analysis.cc:240:33:   required from here
../../include/Rivet/Tools/RivetYODA.hh:108:29: error: âĂŸclass YODA::Histo1DâĂŹ has no member
     named âĂŸsameBinningâĂŹ
    return a->sameBinning(*b);
                             ^
../../include/Rivet/Tools/RivetYODA.hh: In instantiation of âĂŸbool Rivet::bookingCompatible(
     TPtr, TPtr) [with TPtr = std::shared_ptr<YODA::Histo2D>]âĂŹ:
../../include/Rivet/Analysis.hh:1001:56:   required from âĂŸstd::shared_ptr<_Tp1> Rivet::
     Analysis::addOrGetCompatAO(std::shared_ptr<_Tp1>) [with AO = YODA::Histo2D]âĂŹ
Analysis.cc:316:33:   required from here
../../include/Rivet/Tools/RivetYODA.hh:108:29: error: âĂŸclass YODA::Histo2DâĂŹ has no member
     named âĂŸsameBinningâĂŹ
../../include/Rivet/Tools/RivetYODA.hh: In instantiation of âĂŸbool Rivet::bookingCompatible(
     TPtr, TPtr) [with TPtr = std::shared_ptr<YODA::Profile1D>]âĂŹ:
../../include/Rivet/Analysis.hh:1001:56:   required from âĂŸstd::shared_ptr<_Tp1> Rivet::
     Analysis::addOrGetCompatAO(std::shared_ptr<_Tp1>) [with AO = YODA::Profile1D]âĂŹ
Analysis.cc:399:33:   required from here
../../include/Rivet/Tools/RivetYODA.hh:108:29: error: âĂŸclass YODA::Profile1DâĂŹ has no member
     named âĂŸsameBinningâĂŹ
../../include/Rivet/Tools/RivetYODA.hh: In instantiation of âĂŸbool Rivet::bookingCompatible(
     TPtr, TPtr) [with TPtr = std::shared_ptr<YODA::Profile2D>]âĂŹ:
../../include/Rivet/Analysis.hh:1001:56:   required from âĂŸstd::shared_ptr<_Tp1> Rivet::
     Analysis::addOrGetCompatAO(std::shared_ptr<_Tp1>) [with AO = YODA::Profile2D]âĂŹ
Analysis.cc:476:33:   required from here
../../include/Rivet/Tools/RivetYODA.hh:108:29: error: âĂŸclass YODA::Profile2DâĂŹ has no member
     named âĂŸsameBinningâĂŹ
../../include/Rivet/Tools/RivetYODA.hh: In function âĂŸbool Rivet::bookingCompatible(TPtr, TPtr)
      [with TPtr = std::shared_ptr<YODA::Histo1D>]âĂŹ:
../../include/Rivet/Tools/RivetYODA.hh:109:3: warning: control reaches end of non-void function
     [-Wreturn-type]
   }
   ^
../../include/Rivet/Tools/RivetYODA.hh: In function âĂŸbool Rivet::bookingCompatible(TPtr, TPtr)
      [with TPtr = std::shared_ptr<YODA::Histo2D>]âĂŹ:
```

Rivet requests/issues

Rivet requests/issues

Requests:

- See Stefan's talk
- HZTOOL interface (see Simon's talk)
- ROOT for histogramming

Issues:

- Rivet is slow: typically even LEPI jet/event shape analysis can be as slow as 1000 events per second vs. 100000 per second with ROOT or PAW
- Rivet cannot be parallelized – cannot be faster on multicore machines

Other requests to MC related software

- Photon PDFs in LHAPDF6
- All PDFs used in HERA analyses in LHAPDF6
- Proper Fortran interface for LHAPDF6
- Faster LHAPDF6

Outlook

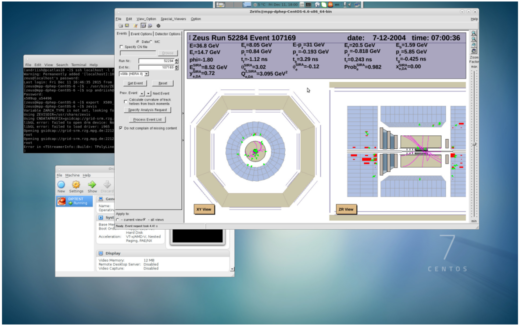
- Both ZEUS and H1 are doing interesting physics with preserved data.
- New collaborators are welcome!
- . . .
- Time for discussion.

Backups

Software environment/VM:ZEUS

A VM image based on SL7 is available

- ZEUS software: ROOT, MC simulation, event display, file catalogue, setup scripts etc.
- Modern MC generators, FastJet, cernlib, PAW, etc.
- Anything you will want to install. . .
- Agree access and download it.



```
[andriish@pcatlas18 ~]$ ssh localhost -l z
Warning: Permanently added '[localhost]:10
zeus@localhost's password:
Last login: Fri Dec 11 16:46:35 2015 from
[zeus@mpp-dphep-CentOS-6 ~]$ . /usr/bin/ZE
[zeus@mpp-dphep-CentOS-6 ~]$ scp andriish@
Password:
x509up_u54496
[zeus@mpp-dphep-CentOS-6 ~]$ export  X509_
[zeus@mpp-dphep-CentOS-6 ~]$ zevis
Variable ZARCH_TYPE is not set, looking fo
Using ZEVISDIR=/usr/share/zevis
Using CNDATAPREFIX=gsidcap://grid-srm.rzg.
libGL error: failed to open drm device: No
libGL error: failed to load driver: i965
Opening gsidcap://grid-srm.rzg.mpg.de:2212
root
Opening gsidcap://grid-srm.rzg.mpg.de:2212
root
Error in <TStreamerInfo::Build>: TPolyLine
)
```

ZeVis@mpp-dphep-CentOS-6.6-x86_64-bin

Zeus Run 52284 Event 107169 date: 7-12-2004 time: 07:00:36

E=36.8 GeV, E_t=8.05 GeV, E-p_z=31 GeV, E_f=20.5 GeV, E_b=1.59 GeV

E_r=14.7 GeV, p_t=0.84 GeV, p_x=-0.193 GeV, p_y=-0.818 GeV, p_z=5.85 GeV

phi=-1.80, t_f=-1.12 ns, t_b=3.29 ns, t_r=0.243 ns, t_g=-0.425 ns

E_e^{SIRA}=8.52 GeV, θ_e^{SIRA}=3.02, ϕ_e^{SIRA}=-0.12, $Prob_e^{SIRA}$=0.982, $x_{e,DA}^{SIRA}$=0.00

$y_{e,DA}^{SIRA}$=0.72, $Q^{2,SIRA}_{e,DA}$=3.095 GeV^2