Update on ZEUS data preservation and usage

Achim Geiser, DESY Hamburg, Germany

for the ZEUS collaboration

4th DPHEP collaboration workshop, CERN, 2. 10. 2024

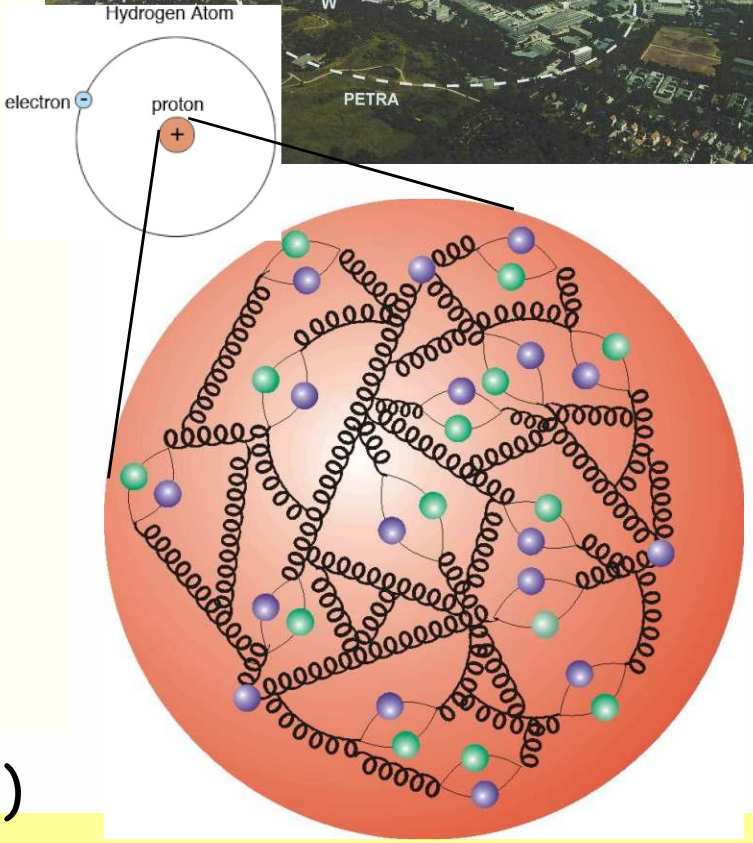
- Reminder: Why? (motivation)
- Reminder: How? (challenges)
- What? (achievements)

What was/is HERA?

- The world's (up to EIC) **unique electron proton collider** with **International Particle Physics Experiments** which recorded high energy electron-proton collisions at DESY in Hamburg, Germany

- **Physics data taking: 1992-2007**

- one of main physics goals:
  measure structure of the proton to
  ~10^{-18} m, i.e. 1/1000 of proton size
  ("X ray" of proton with electrons)
  used e.g. in measurements of **Higgs properties** at LHC



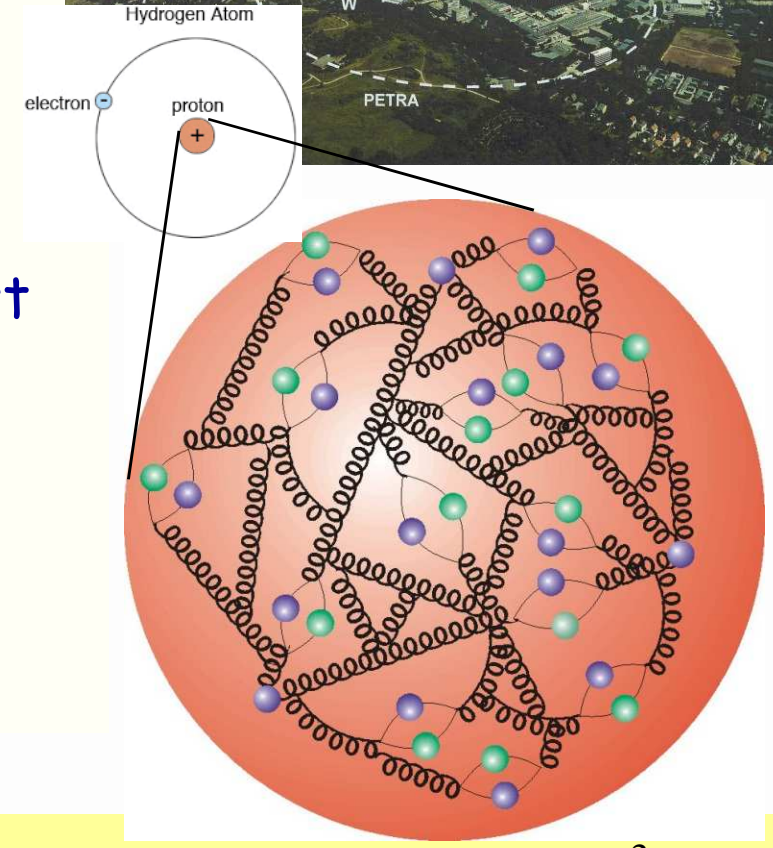
- also well suited to study **general QCD** and electroweak physics + **proton spin** (Hermes)

What is ZEUS ?

- **International Particle Physics Collaboration** which recorded high energy electron-proton collisions with detectors at the world's (so far) unique lepton-proton collider **HERA** at DESY in Hamburg, Germany

- **Physics data taking: 1992-2007**

- General purpose detector suited for almost any physics topic relevant in ep collisions



Motivation is important

- Data preservation should not and cannot be a goal in itself.

- Find good scientific motivations/examples why the data should be preserved and used, and promote them.

- This will enhance the chances that you will get* (part of) the funding needed to do the technical parts of data and knowledge preservation.

* it is always too little

My personal generic vision

formulated many years ago:

with ~1% of additional resources aim to achieve

~10% additional scientific output (e.g. physics papers)

from both external and internal use of **preserved or open data**

over lifetime of experiment/project + 10-20 years

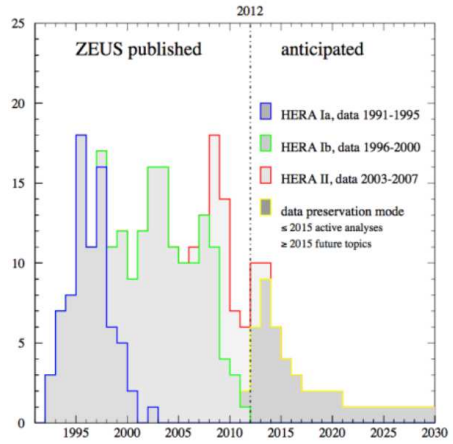
recent extensions in view of e.g. PUNCH4NFDI:

enable common access to and analyses of
HEP, Hadron, Astroparticle and Astrophysical data

-> see presentation tomorrow

ZEUS/HERA data preservation

- Data and knowledge preservation project internally started within ZEUS experiment in 2006 (generalized 2009)



- HERA experiments (incl. ZEUS) + DESY/IT are core co-authors of 2012 DPHEP study group document

- DESY and MPP are co-founding members of Collaboration Agreement for the DPHEP project supported by ICFA (May 2014) (other related institutes have MoUs with DESY)

- workshop on Future Physics with HERA Data at DESY (see backup) (Nov. 2014, end of H1/ZEUS funding) **physics projects steadily being implemented !**

DPHEP data preservation levels

DPHEP = Data Preservation in High Energy Physics arxiv:1205.4667

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses -> education
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data -> raw data

Table 3: Various preservation models, listed in order of increasing complexity.

- ZEUS: level 3 (data and existing Monte Carlo (MC) data), works 'out-of-the box' across system/ROOT updates

  level 4 (additional Monte Carlo data) needs containers

- H1 and HERMES: level 4 (both Raw and reconstructed data)

Challenge: What is the "Data"?

- "Data" = recorded events, simulated events, metadata,
  + related software, knowledge, and documentation

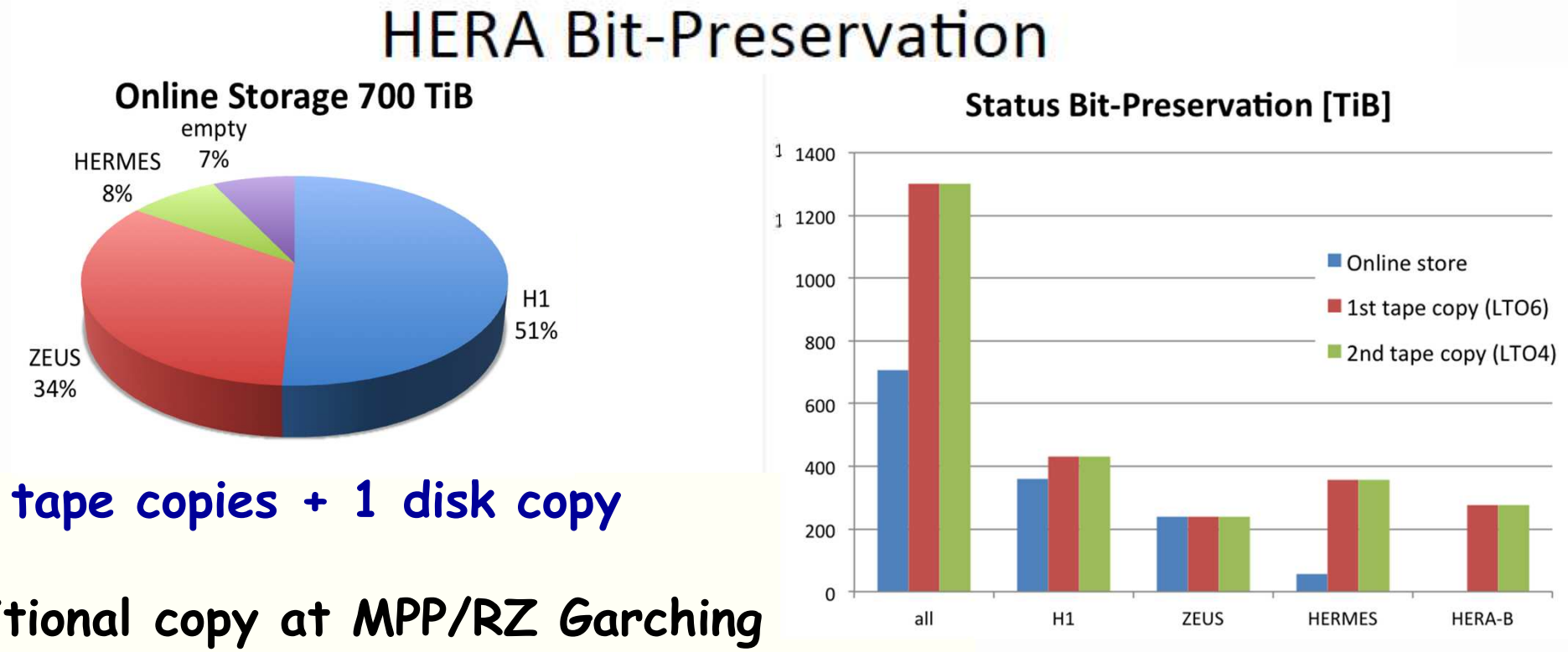
- Bit preservation and data access (computing):
  existing data and MC samples

- Software preservation:
  simulation, reconstruction, analysis, event display

- Documentation:
  analog and digital archives, web pages (see backup)

Challenge: Bit preservation

- at DESY: common approach for all three HERA experiments



2 tape copies + 1 disk copy

+ additional copy at MPP/RZ Garching (for ZEUS part)

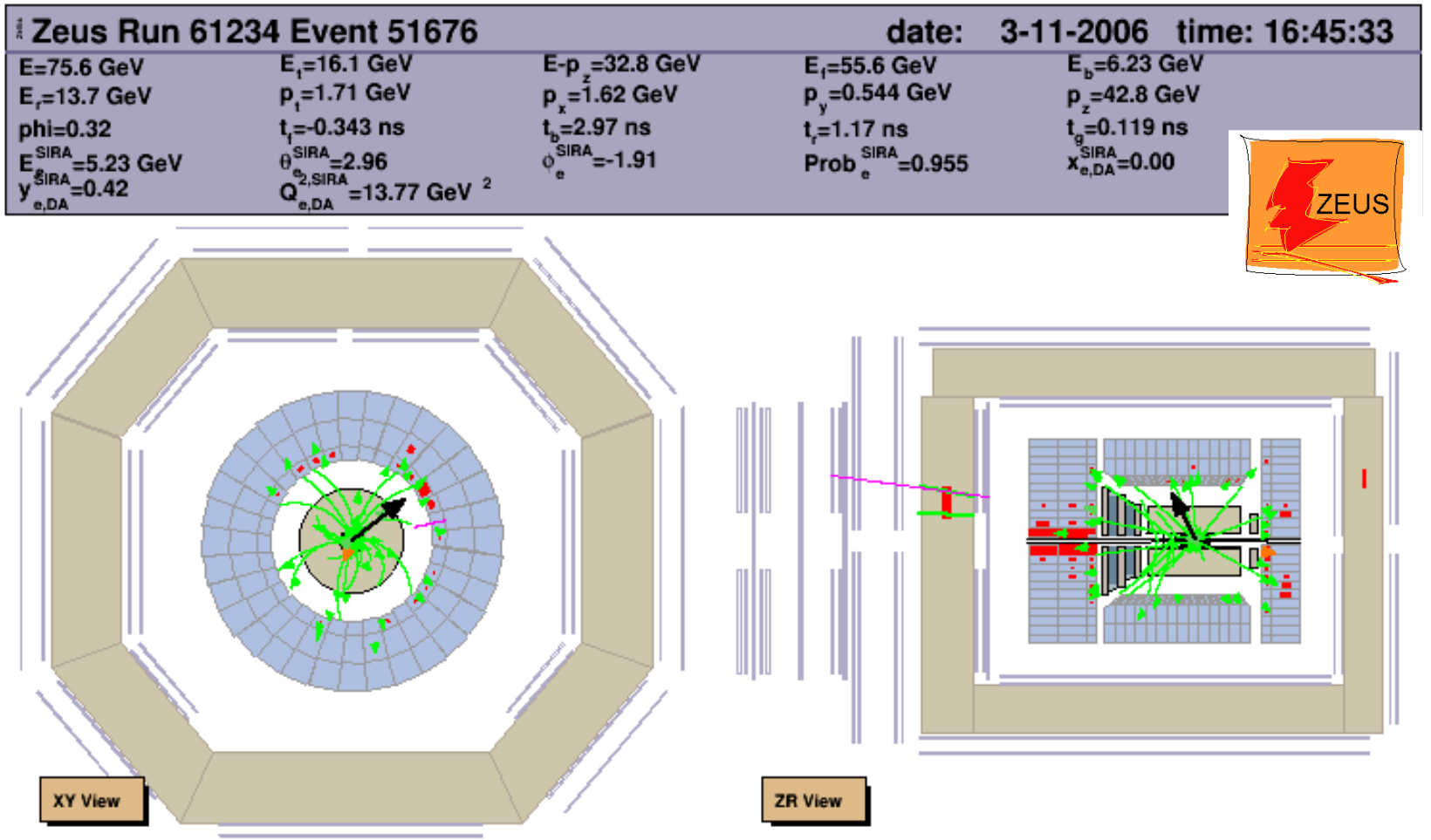
Challenge: Computing

- all remaining dedicated hardware for all three HERA experiments decommissioned since 2014/15.
- **long term data access guaranteed by DESY IT**.

- currently access to preserved data at DESY on generic "BIRD" batch farm (National Analysis Facility, NAF), e.g. ~15 ZEUS users (integrated).
- shared opportunistically with LHC and other experiments but fully sufficient for relatively modest HERA needs.
- job submission via dedicated servers (EL9) maintained by DESY IT. Can also be used for interactive debugging and event display.

- access to ZEUS data also at MPP Munich

What do ZEUS data look like?



event display

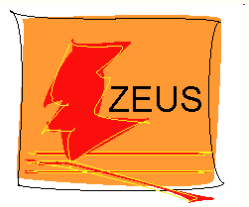
from ZEUS

"Common Ntuple"

complicated physics data content: for useful analysis, need

significant expert knowledge + documentation + guidance how to use it

Software preservation



- **ZEUS: starting 2006,** unmaintainable (person power) software from 1990's completely replaced by **simplified ROOT common ntuple approach** for analysis; **"flat" root ntuples suited for any physics.**
- **SL5 (2012)-> SL6 -> SL7/EL7 -> EL9 (2024) "transparent".**
- **No porting needed ☺.**

  includes standard MC samples. See *J.Phys.Conf.Ser.* 396 (2012) 022033

- **Virtualization/containerization approach based on frozen SL5 executables** (MPP) for new MC.

  see arXiv:1607.01898

advertised and used since 2012

Common Ntuple analysis model

- **ZEUS Common Ntuple:** **Motto: keep it simple!**

  flat (simple) ROOT-based ntuple (same format as PAW ntuple converted with h2root) containing high level objects (electrons, muons, jets, energy flow objects, …) as well as low level objects (tracks, CAL cells,

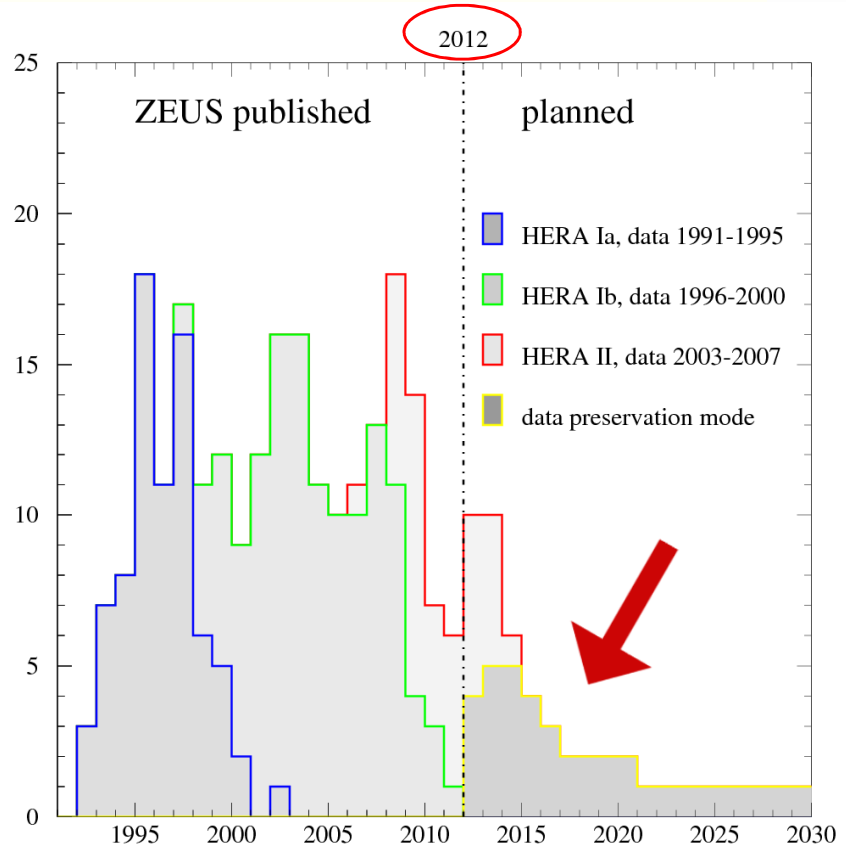
- **Well tested !**

  all recent ZEUS physics papers based on Common Ntuples

- **Easy to use**

  several recent ZEUS results based on results produced by Master students.

  PhD students can produce a ZEUS paper within only a fraction of their PhD time (e.g. ~6 months -1 year)



status end 2024

ZEUS physics papers

majority of ZEUS papers produced

in "data preservation mode"

already since 2012 (31 papers)

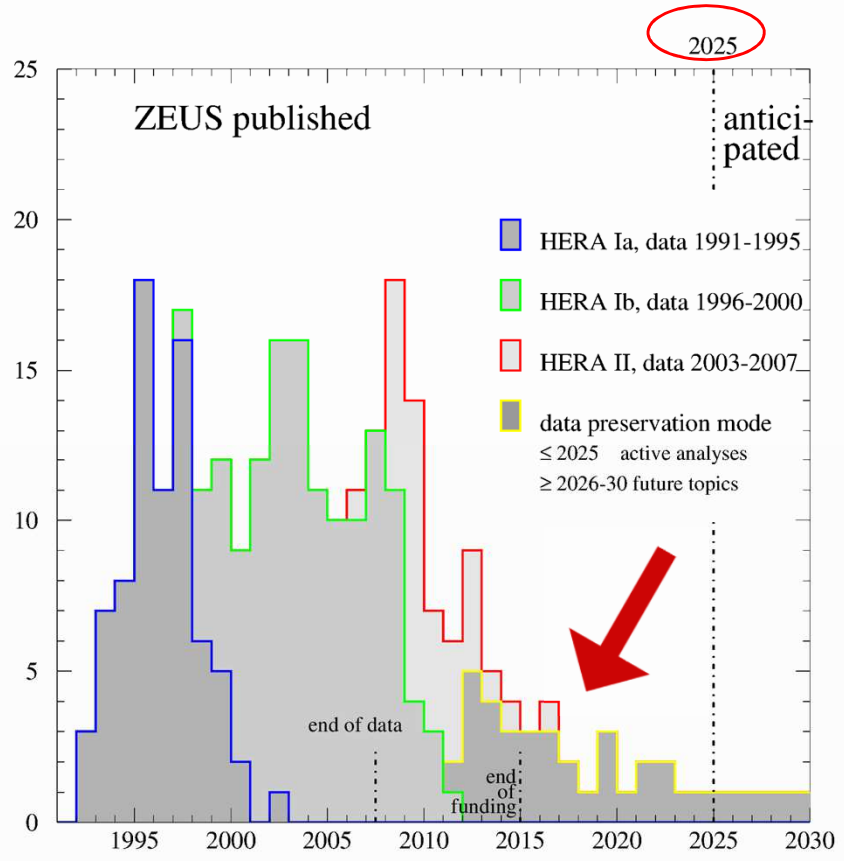
long term: ~1-2 paper/year -> >~2030

expect ~10% of total ZEUS output

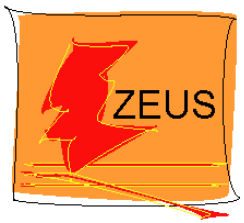
~80% of these would never exist
without dedicated data preservation

How to get access to data? See backup.

PAW setup for this plot preserved ☺



Recent example physics papers



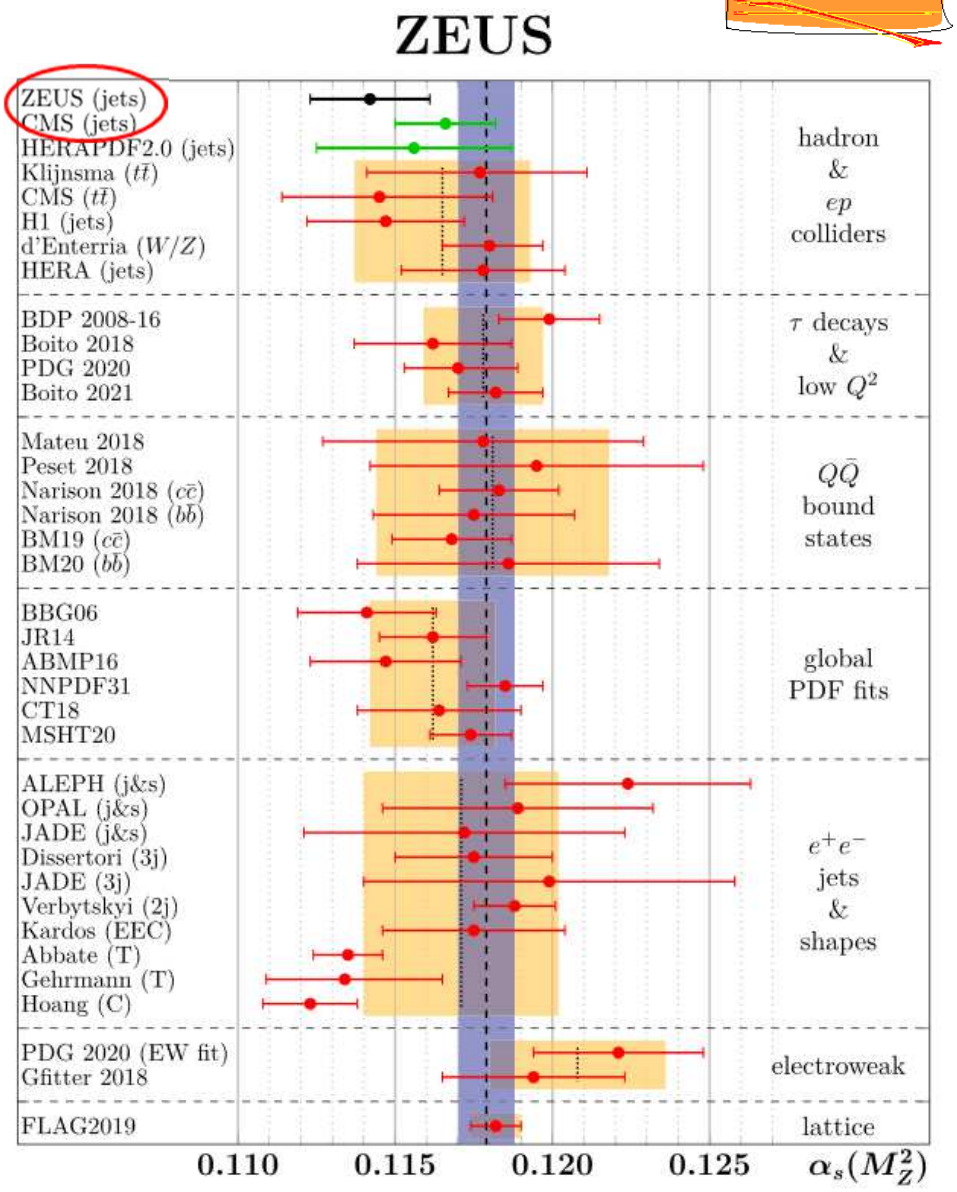
Eur. Phys. J.C 83 (2023) 11, 1082
arXiv:2309.02889

- **measurements of strong coupling constant at NNLO**

  (QCD equivalent of QED Sommerfeld constant)

  from HERA jets

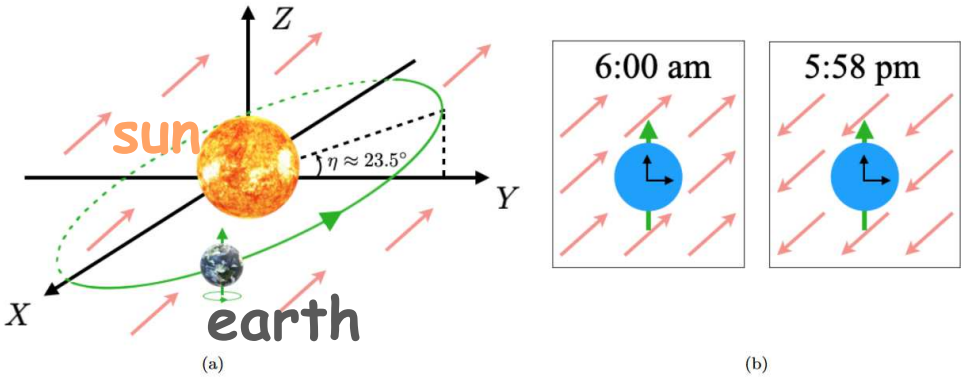
- **world competitive**

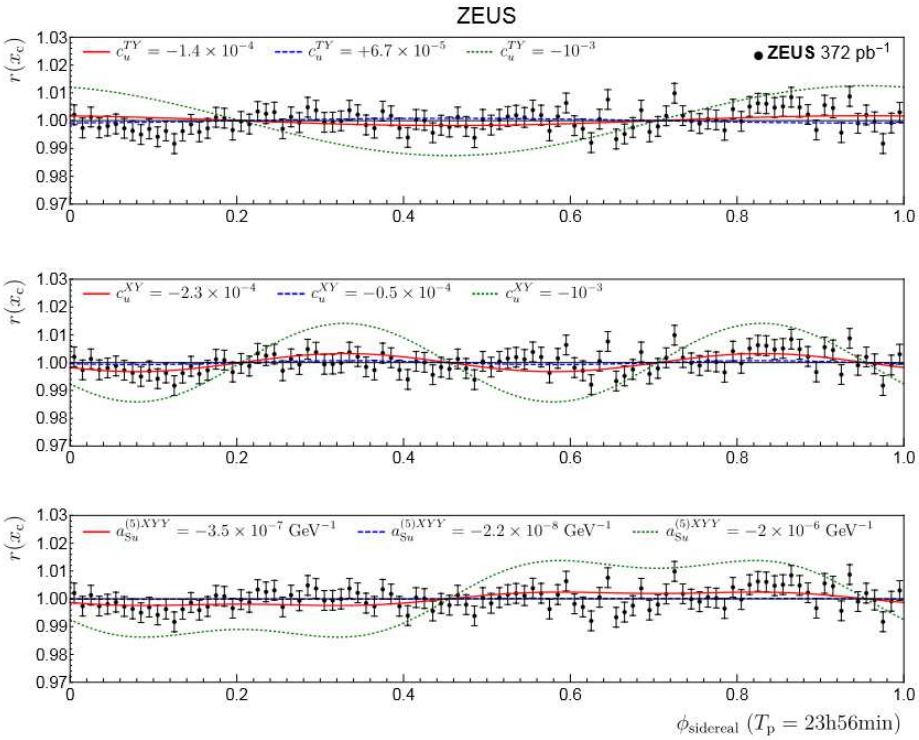


Tests of Lorentz invariance

Phys. Rev. D 107 (2023) 9, 092008
arXiv:2212.12750

sidereal (~daily) variations





- **World best limits for effective Lorentz invariance violation from EFT coupling of quarks to some unknown cosmic background field**
- **ZEUS data analysis done by theorists ! ☺**

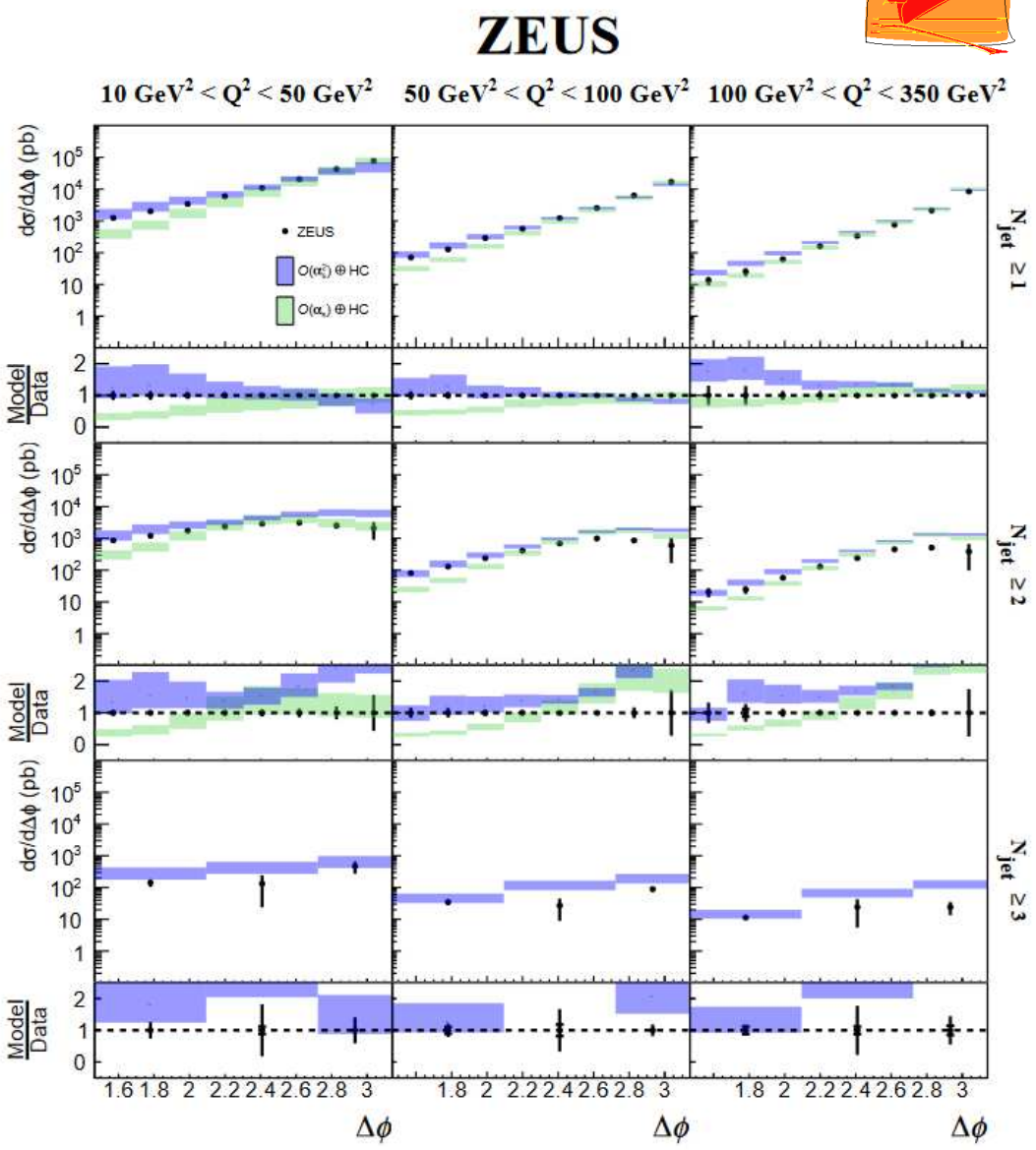
Jet/lepton azimuthal correlations in DIS

arXiv:2406.01430

- **QCD works successfully for jets down to p_T > 2.5 GeV !**

  both theory calculations and MC

- **ZEUS analysis done in collaboration with EIC community**



Conclusions and Outlook

- HERA data are scientifically unique and worth preserving !
- ZEUS data preservation program is a success !
  large parts of original ZEUS data preservation plan successfully implemented
- 17 years after end of data taking in 2007, thanks to data and knowledge preservation, ZEUS scientific output continues at a significant rate, for very little cost
  (a tiny bit of „official" funding would help to do even better)
- about 30% of total number of ZEUS papers produced after end of data taking. Made possible through substantial support by collaborators, host lab (DESY, IT), and external institutes!
- expect ~10% of total scientific output to originate from data preservation efforts (i.e. after end of funding), if long term sustainability is continued
  (large part of that already done!)
- Bottleneck: long term "visible" person power

Backup

Size of data sets

Root files (officially preserved) units: Tb (status 4.9.13, still valid)

HERA II	v02	v06	v08	HERA I	v08 +v07	total
Data	1.9	5.2	7.0	1.7+1.		17.
MC	10.5	64.0	70.	4.8+4.		153.

+30 for future MC

~ **100 million inclusive DIS events** (Q^2>5 GeV^2, triggered almost bias-free)

~ **100 million semi-inclusive photoproduction events** (mainly via p_T>4 GeV dijet trigger)

smaller sets of more specialised triggers/samples (e.g. **heavy flavors, vector mesons**, …)

~ equal sample sizes for e+, e-, righthanded/lefthanded **polarisation**

~ **4 billion MC events**, for almost any analysis

generation of additional MC samples possible (via MPI)

can technically read/analyze full ZEUS data set on one CPU within ~1 day

(for even faster access, many analyzers produce their own mini-ntuples for analysis)

ZEUS software approach

- original ZEUS data format and core software from 1990's
- maintenance of software, simulation and analysis framework needed **~4 FTE/year** (experiment) **+ IT**
- e.g. porting from SL4 to SL5 took about 2 years

-> not sustainable long term

-> go for simplified ZEUS data format:

"Common Ntuples" = flat ROOT ntuples

- almost no dedicated software maintenance needed

-> for new simulation: freeze software and run compiled executables in virtualized environment

- see also https://wwwzeus.mpp.mpg.de

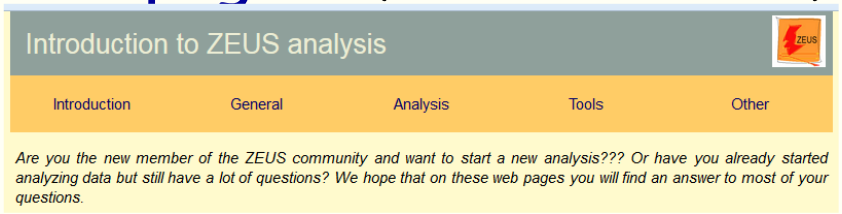
managed at MPP

Analog and digital archive

- full analog archive in DESY library, partially digitized (HERMES)

- all ZEUS technical notes digitized on INSPIRE (via DESY library)
- plain html documentation web pages (DESY web office)
  - ZEUS since 2014

    meeting management -> Indico



  - H1 public web server now also in plain html mode

    Many H1 collaborative tools based on cgi-scripts for accessing oracle.
    Work-around: for critical tools -> local web-server using port 8080 which is not reachable outside firewall.
    Longer term: have to seek for another solution.

  - HERMES web server: on wikimedia, some old cgi scripts hosted on virtual machine
- knowledge preservation also in "human neural networks" (collaboration members)

Workshop:

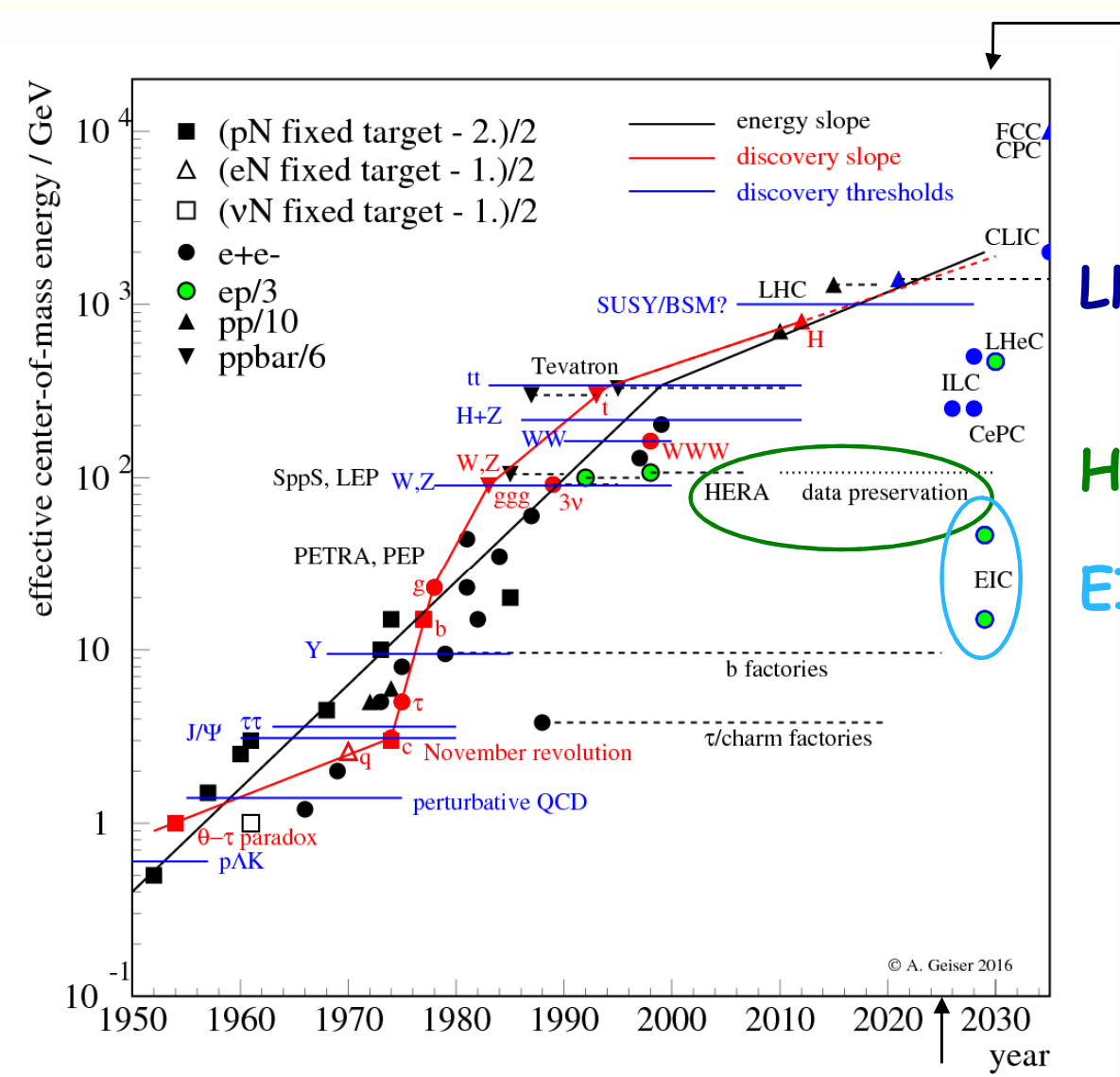
- What do the HERA data still have to say and how are they relevant to other facilities?

- two days with lively discussions and almost 30 presentations
https://indico.desy.de/event/futurehera

- ~ 70 participants, both experimentalists and theorists from across the globe

- -> list of dozens of subjects that are still to be investigated or exploited fully, using the preserved data sets (proceedings in arXiv:1601.01499, arXiv:1512.03624)

Why to preserve/analyze HERA data?

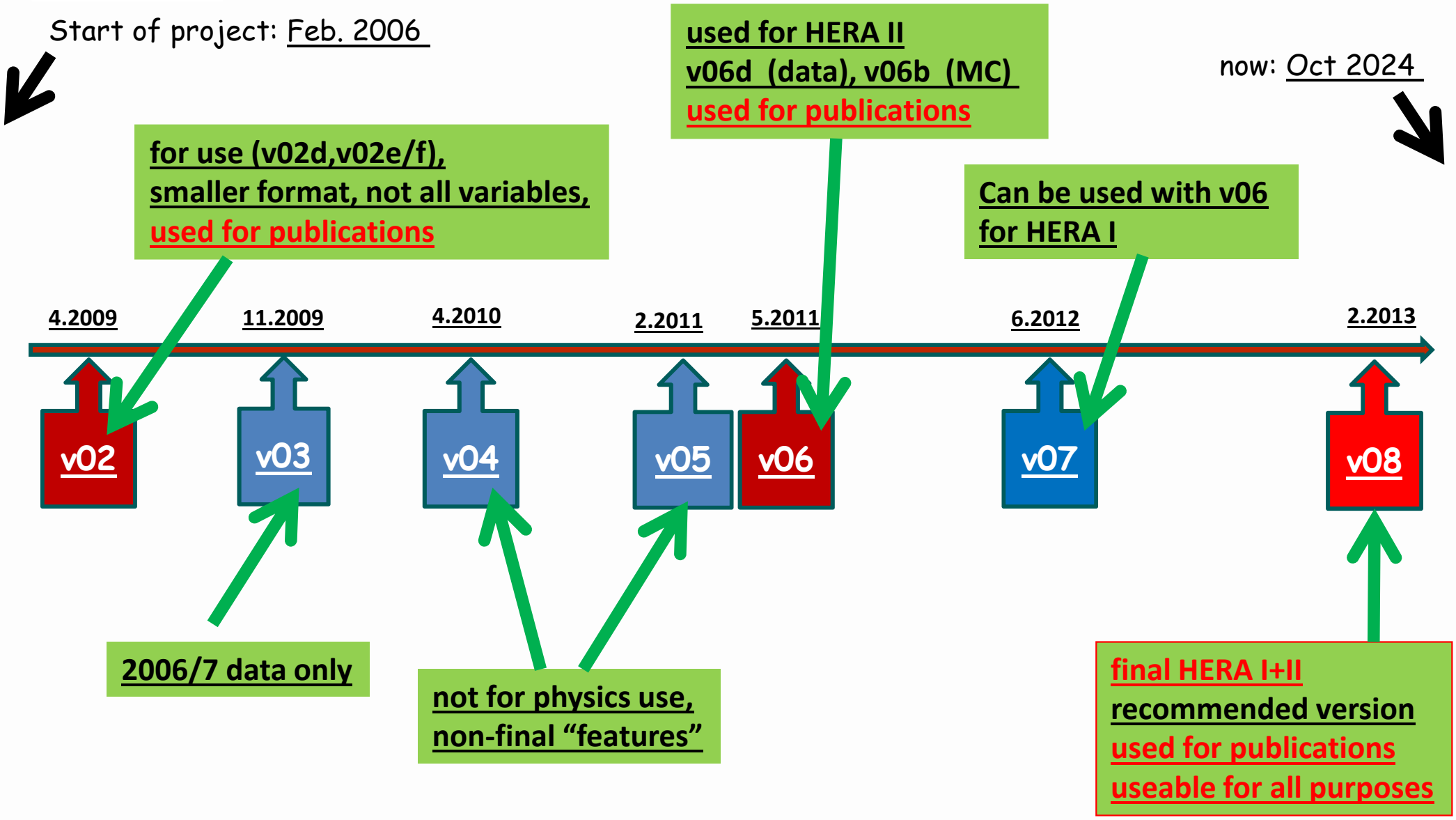


LHC ongoing
and complementary!

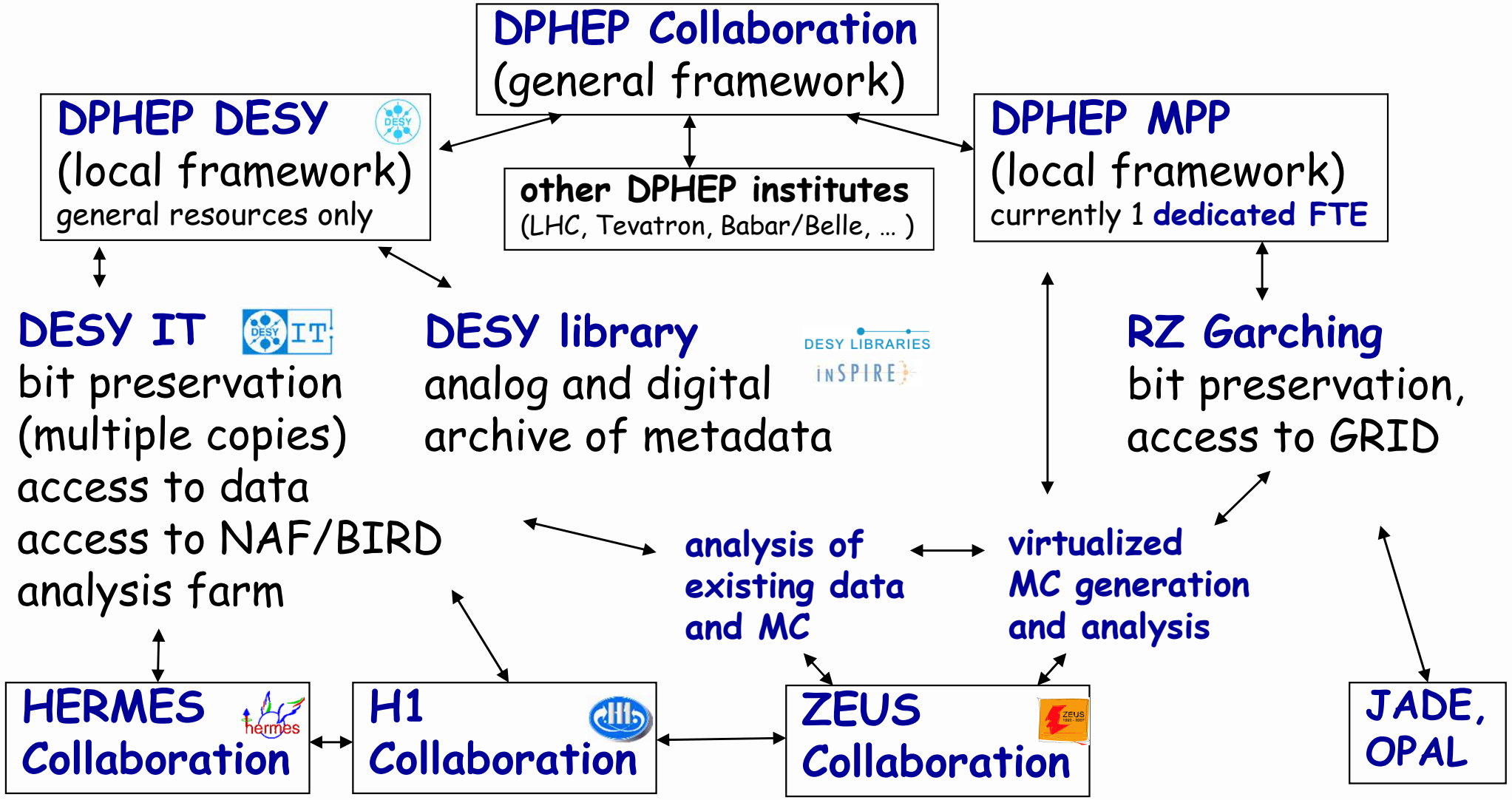
HERA data are unique!

EIC new and complementary!
use synergy!

Available Common Ntuples



HERA Data Preservation Challenge: How to organize the Management?



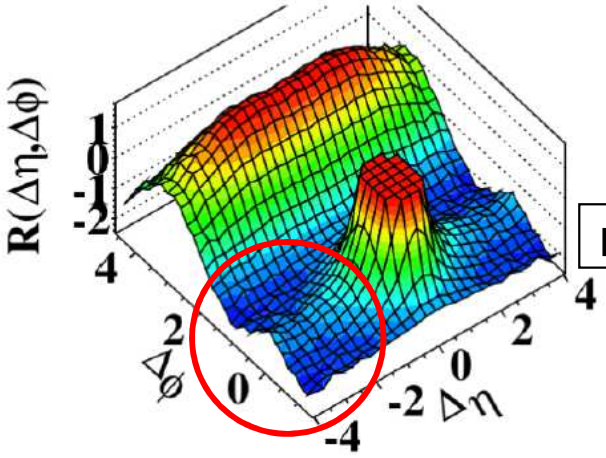
example candidate for cross-experiment archived/open data analysis: "Ridge" in long range particle correlations

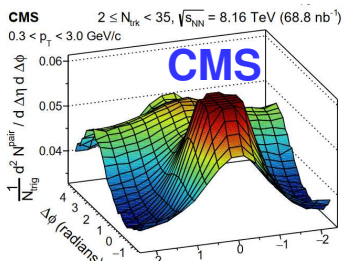
unexpected „Ridge" observed in CMS 2010 pp data

CMS paper

JHEP 1009 (2010) 091

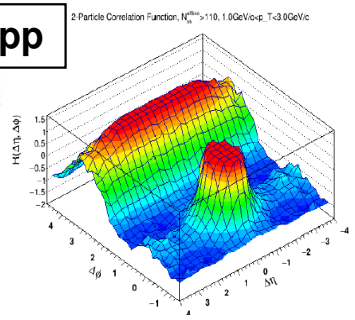
(d) CMS N ≥ 110, 1.0GeV/c<p_T<3.0GeV/c

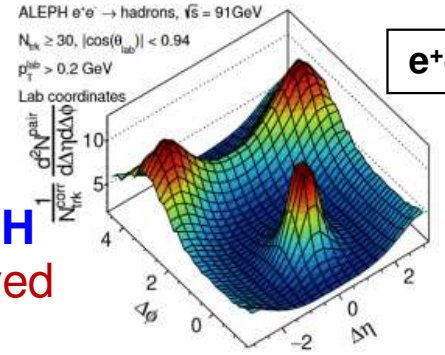




CMS Open Data
(summer student on office desktop)

pp

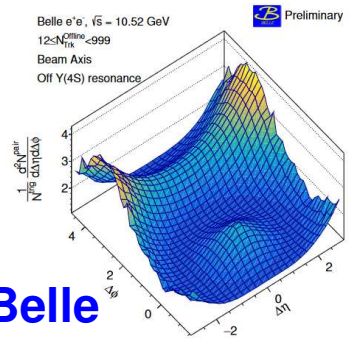




e⁺e⁻

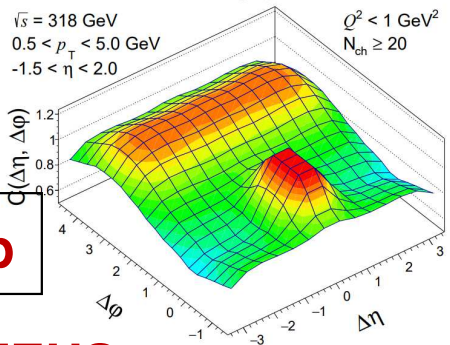
ALEPH archived data

Phys. Rev. Lett. 123 (2019) 212002



Belle

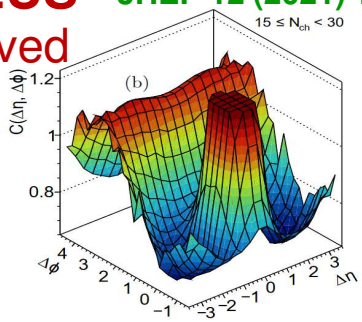
ZEUS



γp

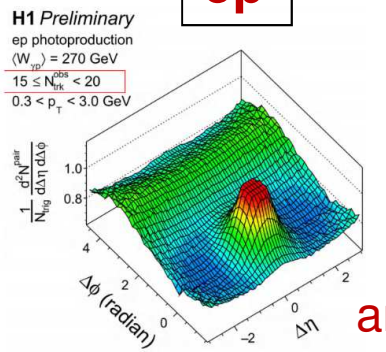
ZEUS archived data

JHEP 12 (2021) 102



JHEP 04 (2020) 070

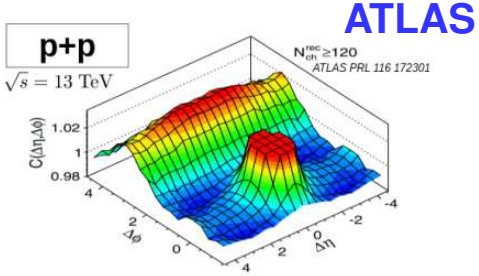
ep

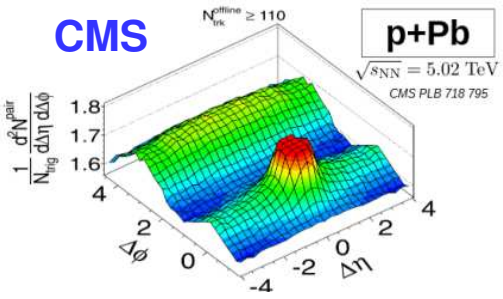


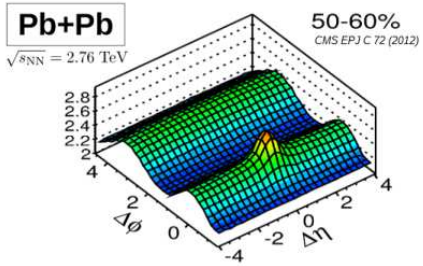
H1 archived data

not complete!

ATLAS







CMS Open Data available

see talks later in workshop

How to get access to the HERA data

ZEUS: (common ntuples, flat root ntuples, only software needed: plain root, almost any version); both HERA I and HERA II data

contact Katarzyna.Wichmann@desy.de (ZEUS spokesperson)

(or me) options:

- either access for specific single project/paper for common publication, or
- become full ZEUS member (no fees/chores beyond working on the physics) and participate in all papers

H1: (dedicated OO framework)

contact Stefan.Schmitt@desy.de (H1 spokesperson)

to become H1 member (no fees fees/chores beyond working on the physics)

HERMES: contact Gunar.Schnell@desy.de (HERMES spokesperson)

"Discoverability"

DPHEP portal:

- http://hep-project-dphep-portal.web.cern.ch

ZEUS web page:

- http://www-zeus.desy.de/

information on ZEUS far from perfect

(**person power** …, in case of availability conflict, content/useability takes preference over (organisation of) documentation)

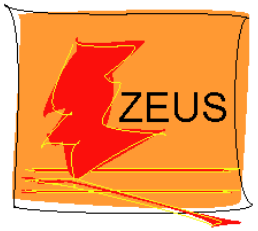
… but we are proud of what we achieved ☺

see also presentation A. Verbytskyi at DIS2016 conference
https://indico.desy.de/contributionDisplay.py?contribId=176&sessionId=7&confId=12482

and ZEUS MPI web page https://wwwzeus.mpp.mpg.de/

How to analyze ZEUS data at DESY?

(additional possibilities at MPI)



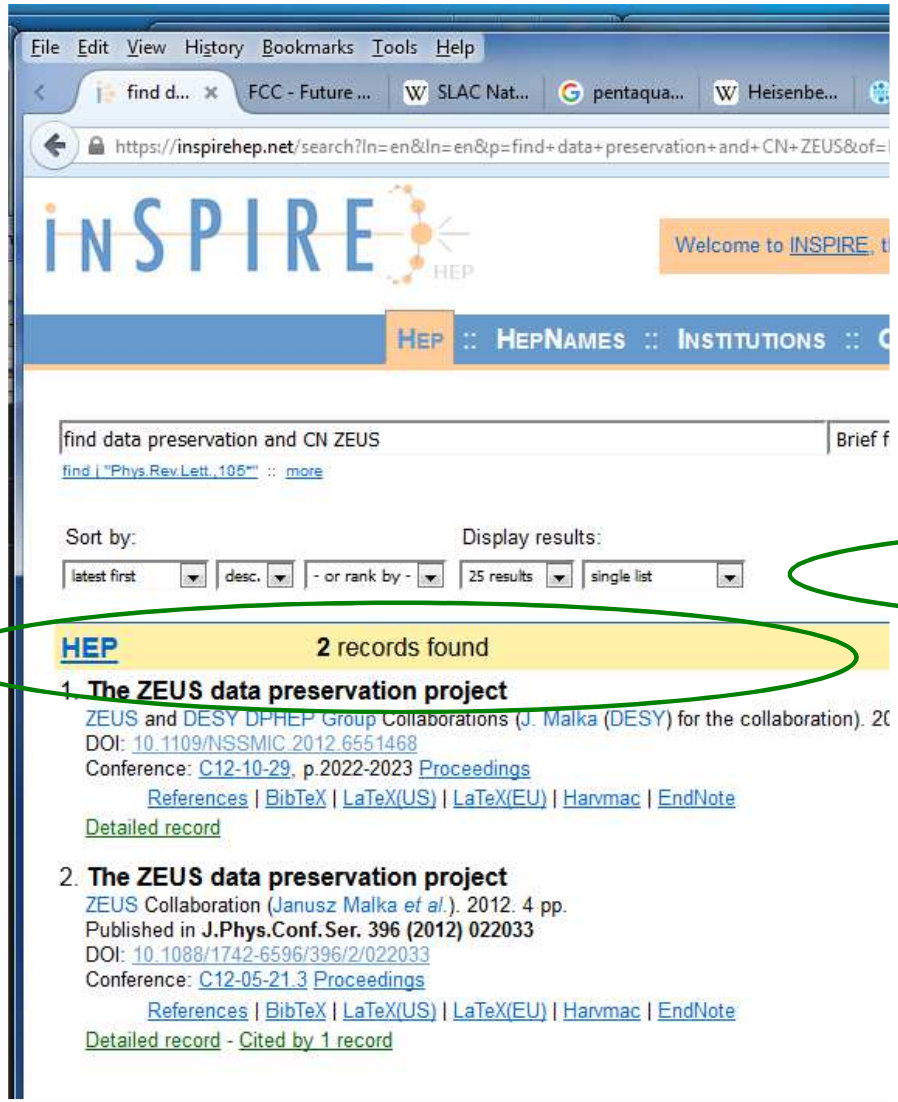
- need:
  - interest in some physics topic ☺
  - agreement with ZEUS management and DESY to obtain
  - ZEUS user account at DESY
    -> access to NAF/BIRD analysis farm via
    ZEUS NAF server (can log on from remote)
  - basic knowledge of generic HEP ROOT package
    (no special ZEUS software to learn!)
  - basic knowledge of particle physics

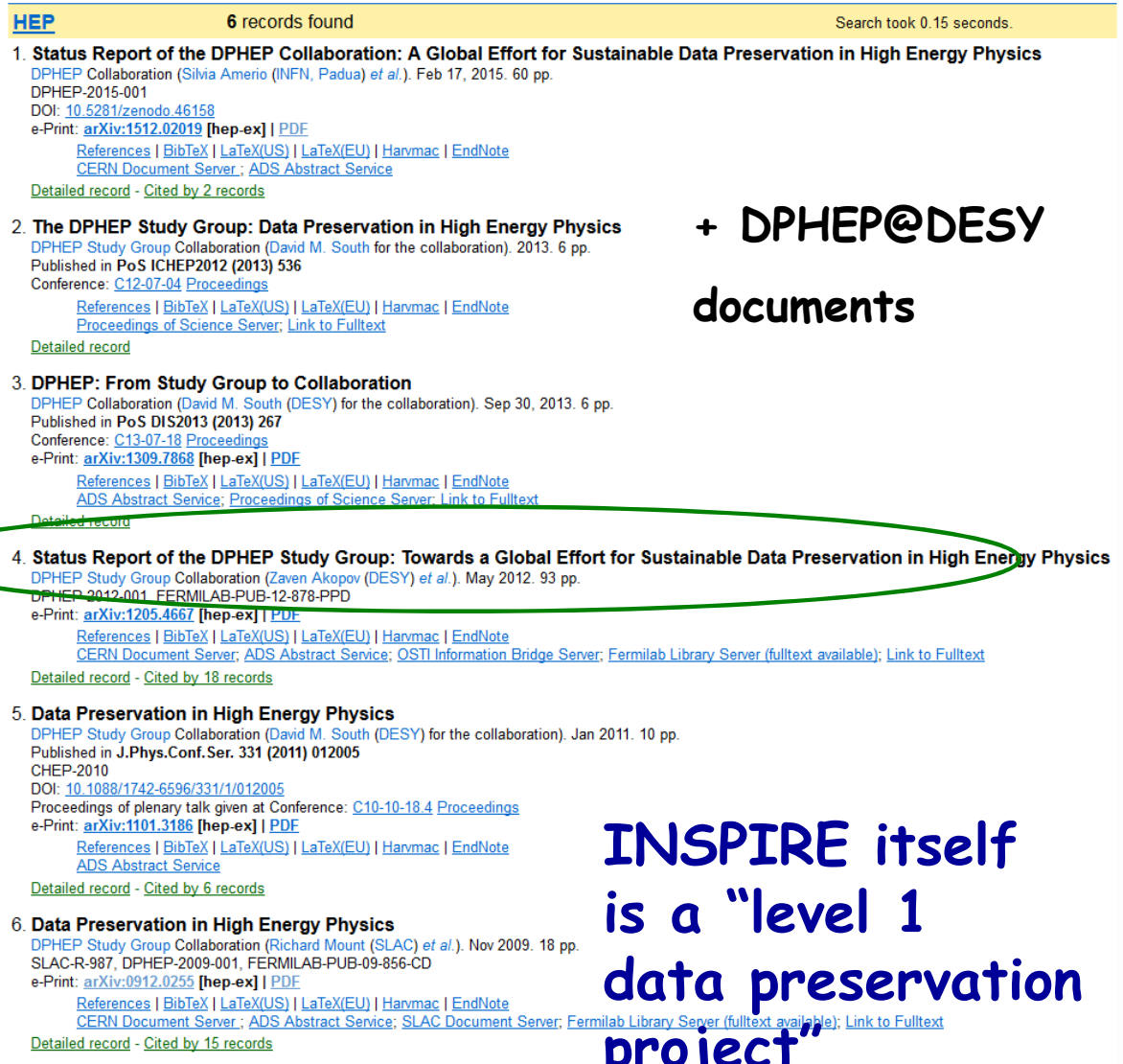
HERA Open Data?

ZEUS might be willing to make (initially part of) its data publicly available, if appropriate nonnegligible temporary person power for proper documentation and curation can be found/paid (no resources within ZEUS).

Any interest from the community?

Publicly available information on DPHEP and ZEUS data preservation





+ DPHEP@DESY documents

INSPIRE itself is a "level 1 data preservation project"