Mapping DMPs to a "past" experiment: the ZEUS experience

Achim Geiser, DESY Hamburg, Germany

ADMP workshop
CERN, Geneva, 30. 6. 2016

- from ZEUS collaboration perspective
- Why? (motivation)
- How? (challenges)
- What? (achievements)

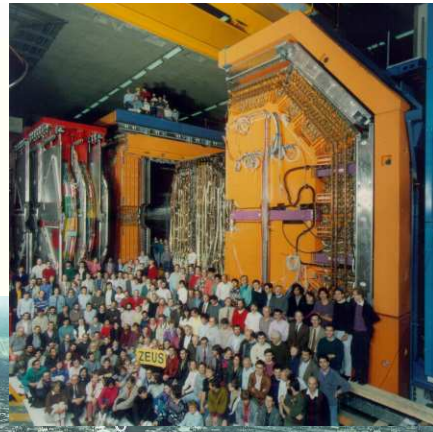
What is ZEUS?

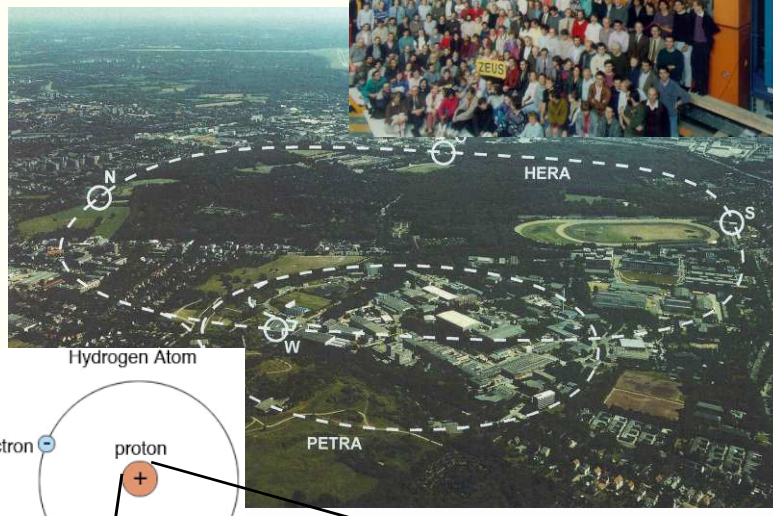
- **International Particle Physics Experiment** which recorded high energy electron-proton collisions at the world's (so far) unique lepton-proton collider **HERA** at DESY in Hamburg, Germany

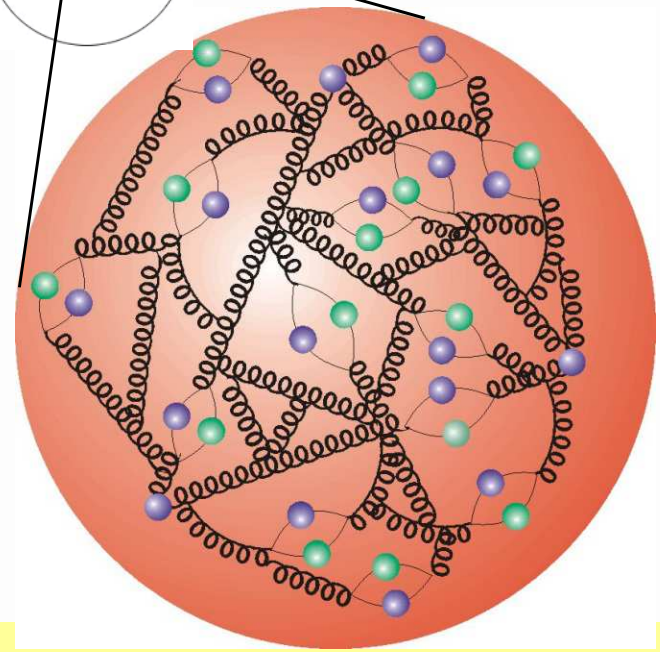
- **Physics data taking: 1992-2007**

- 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)

- also directly useful for LHC, e.g. to measure Higgs properties







Data management in particle physics

- general CERN/LHC and LGC Grid computing perspective addressed by Jamie Sheers (Tuesday)
- CMS point of view addressed in previous talk by Kati Lassila-Perini
- this talk: example perspective on data preservation from a past non-CERN particle physics experiment

  (no claim of general validity)
- for more technical details, see also presentation A. Verbytskyi at DIS2016 conference https://indico.desy.de/contributionDisplay.py?contribId=176&sessionId=7&confId=12482

Why this talk at this meeting?

this meeting is about plans for

Active Data Management !

actual ZEUS experience (also other experiments):

- passive data management (just storing the data somewhere) will not work long term,

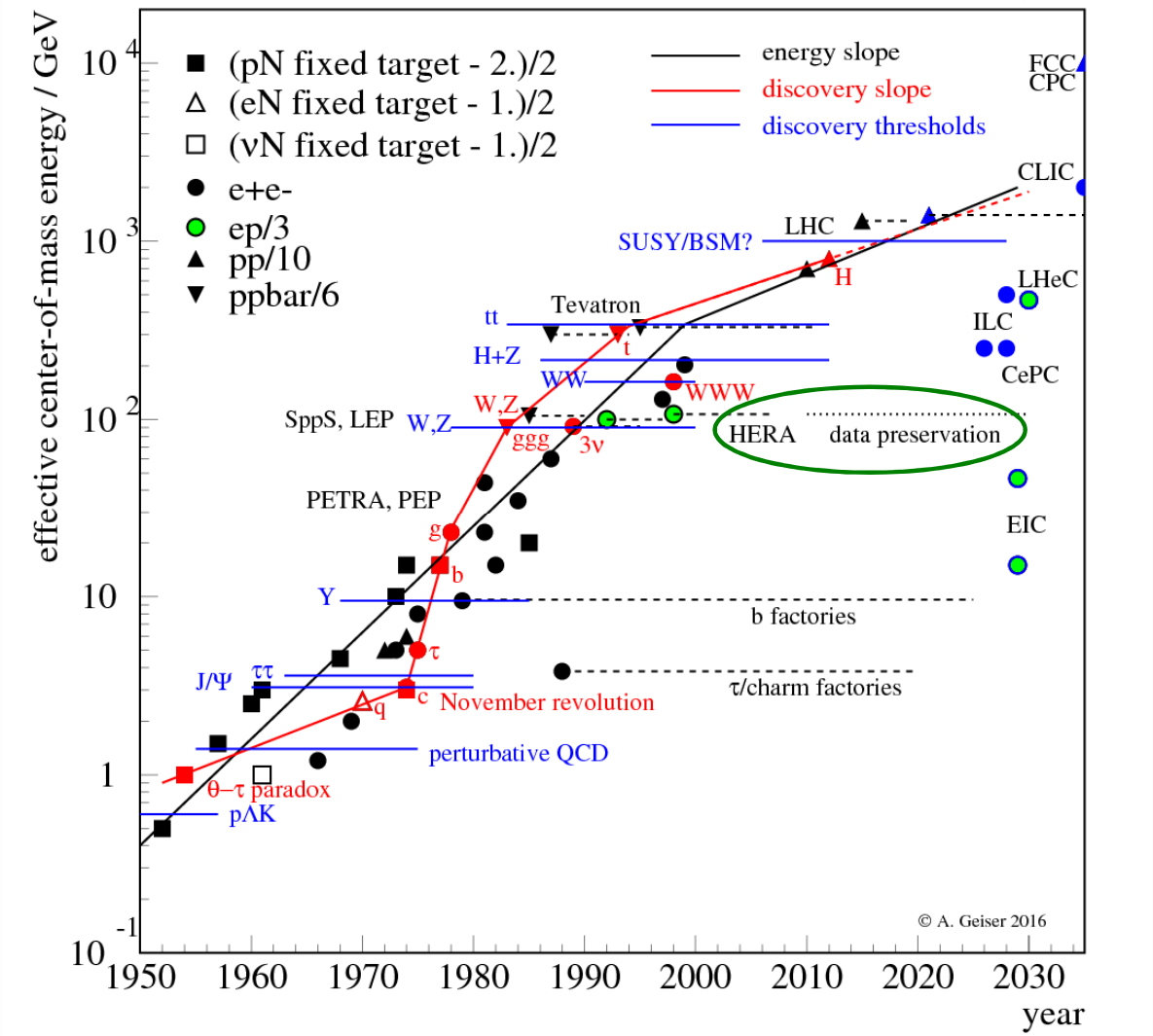
  Active data management is crucial

- **Data** must include metadata, and preservation of software, knowledge, and useability

- **Management** must include manpower needed for long term management, both at IT and user level

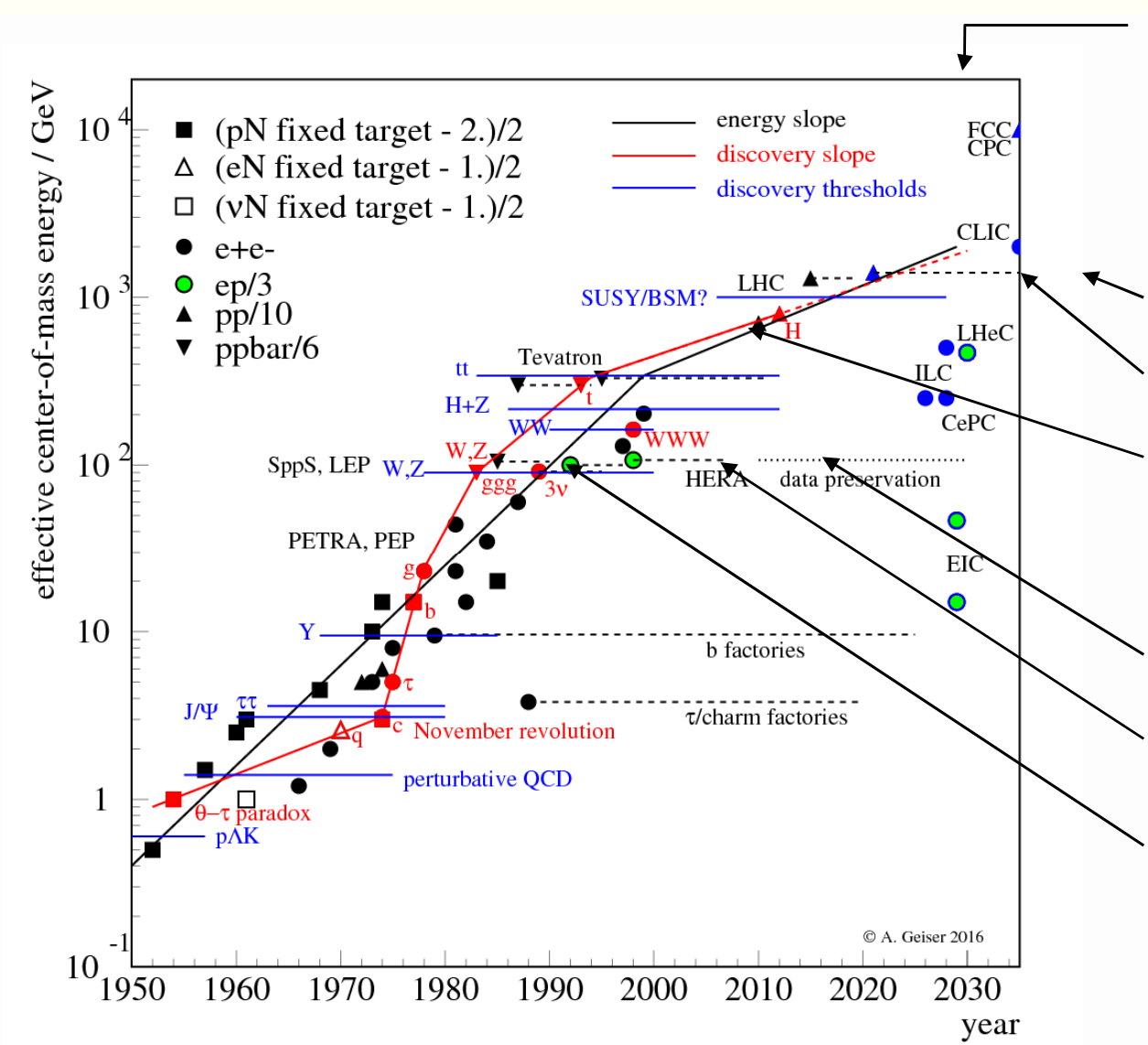
Why to preserve HERA data?

planned new projects



HERA data are unique!

Why to preserve HERA data?



planned new projects

equivalent LHC future

possible end of LHC

start of LHC data taking

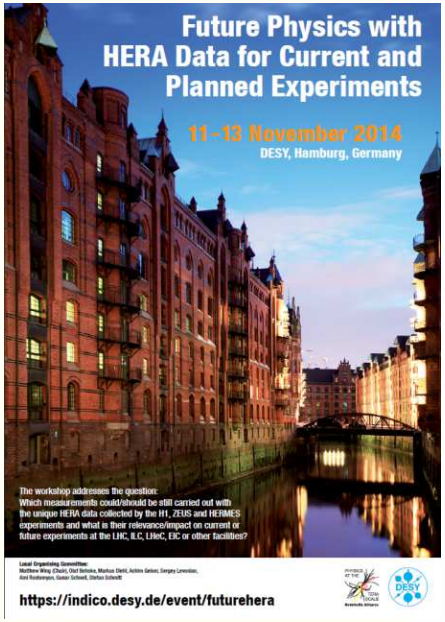
now (HERA data preservation)

end of HERA data taking

start of HERA data taking

Synergy with current experiment: LHC

- LHC collides protons on protons
- detailed knowledge of proton structure is crucial for many LHC physics topics, e.g. for measurement of Higgs boson properties
- in general, many common physics topics

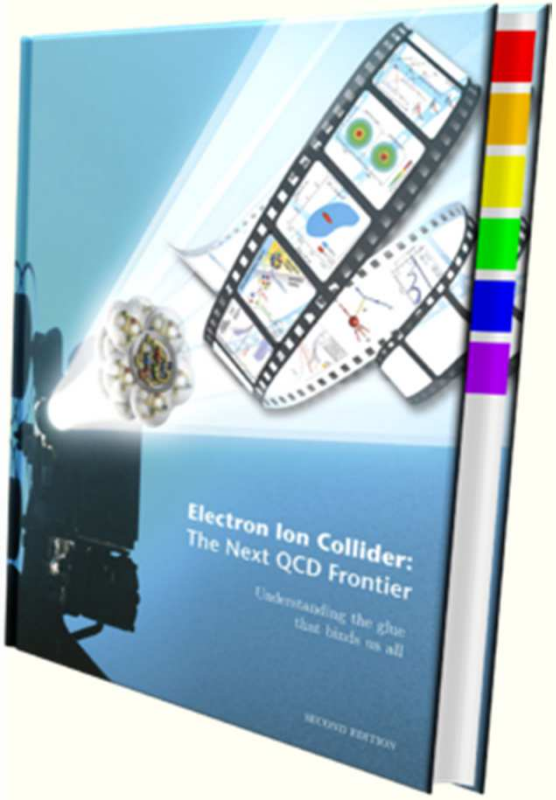


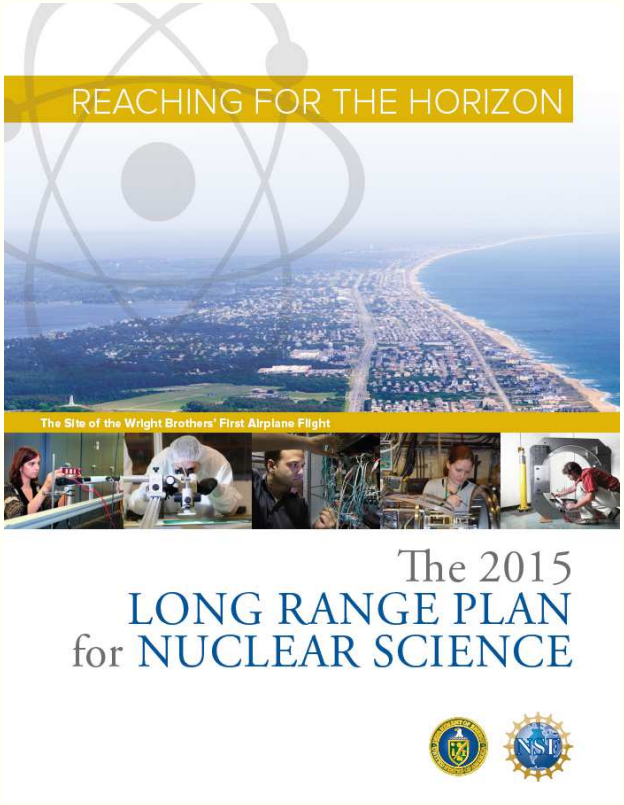
see also

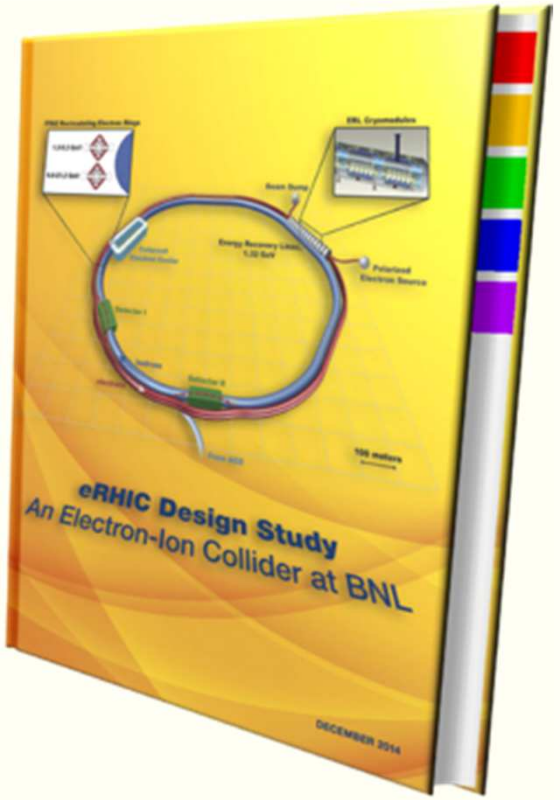
- HERA-LHC workshops, DESY and CERN
- workshop on Future Analysis of HERA data, DESY, November 2014, https://indico.desy.de/conferenceDisplay.py?confId=10523

Synergy with future experiment: EIC

- many EIC topics common with HERA

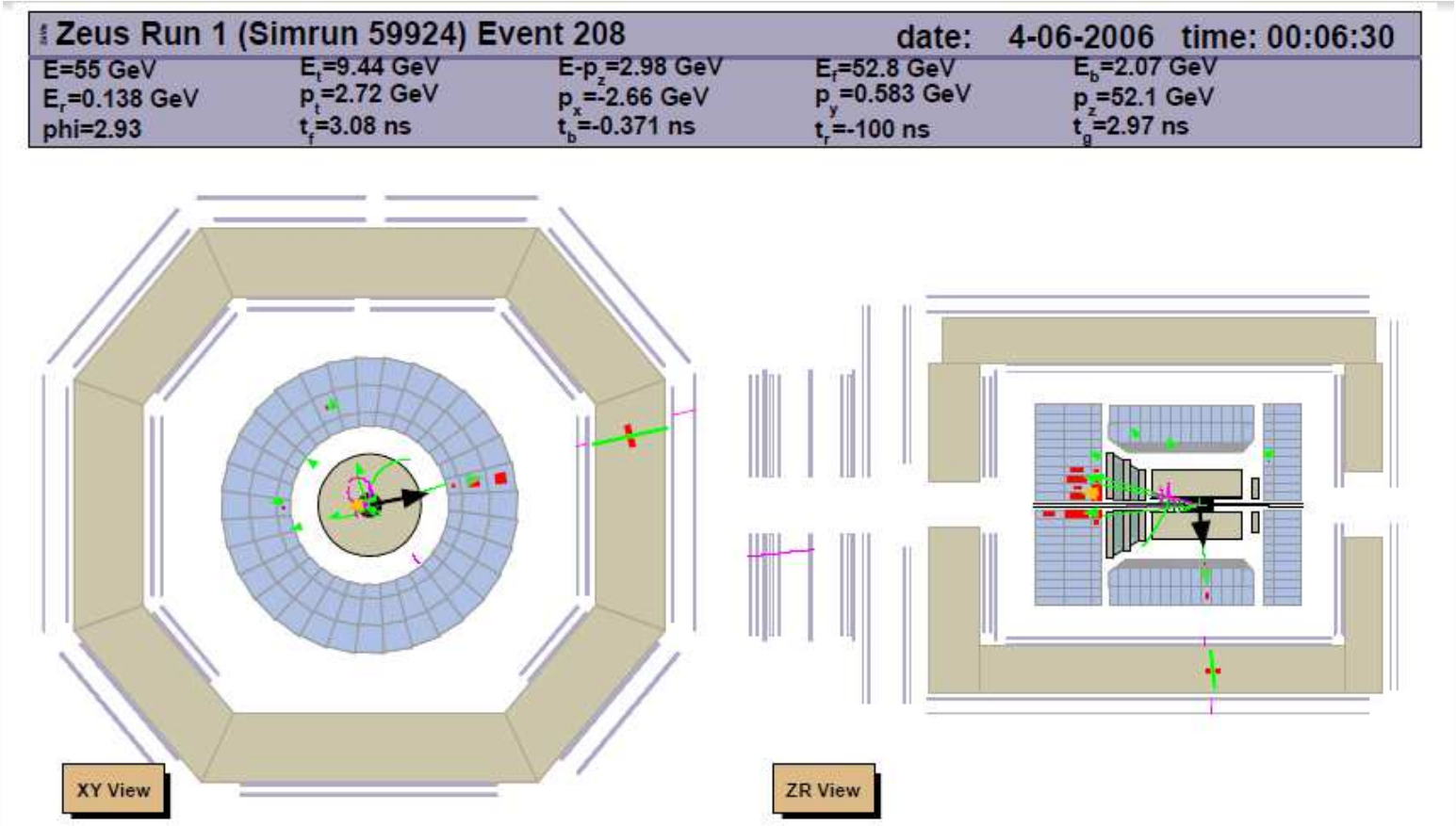






- informal discussions with EIC members on possible common analyses of HERA data started

What do ZEUS data look like?



event display from "Common Ntuple"

complicated data format and content: for useful analysis, need

significant expert knowledge + documentation + guidance how to use it

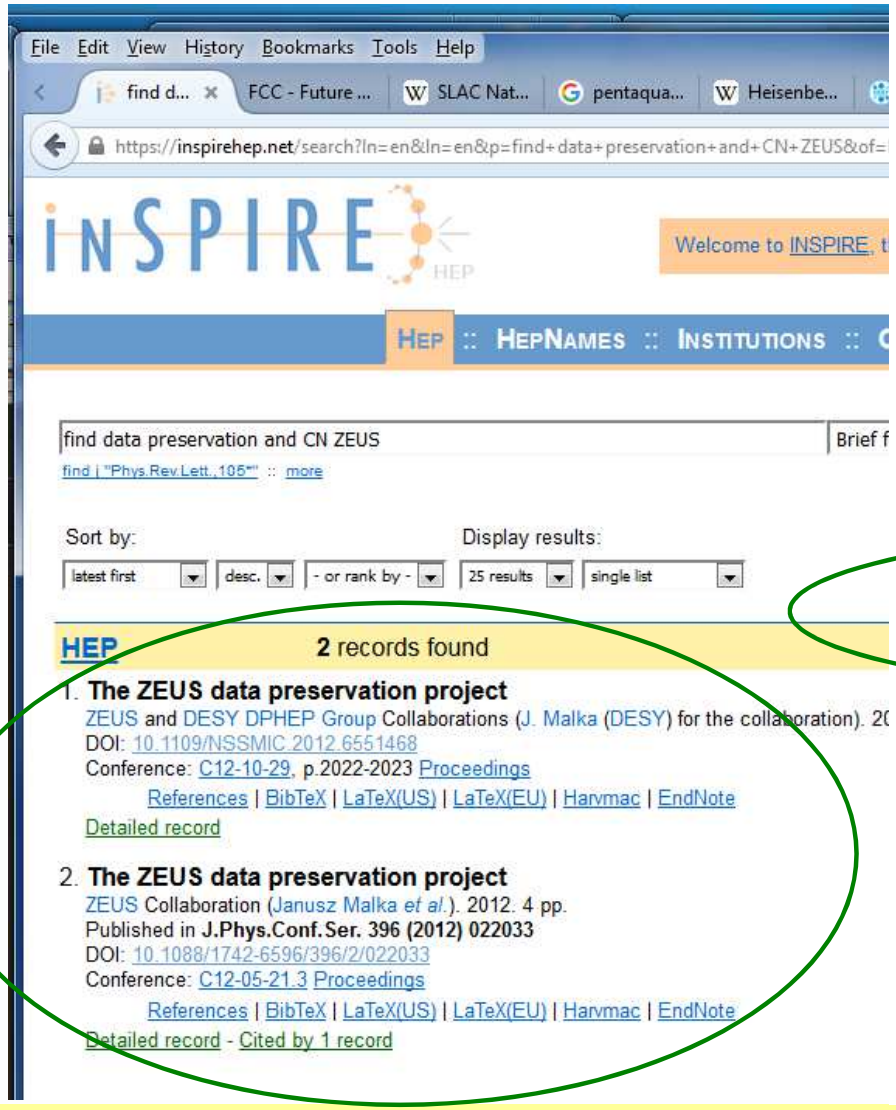
DPHEP data preservation levels

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

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

- ZEUS: level 3 (data and existing Monte Carlo (MC) data),
  level 4 (additional Monte Carlo data)
- other HERA experiments: level 4

Publicly available information on DPHEP and ZEUS data preservation



HEP 6 records found Search took 0.15 seconds.

1. **Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics**
DPHEP Collaboration (Silvia Amerio (INFN, Padua) *et al.*). Feb 17, 2015. 60 pp.
DPHEP-2015-001
DOI: 10.5281/zenodo.46158
e-Print: arXiv:1512.02019 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server ; ADS Abstract Service
Detailed record - Cited by 2 records

2. **The DPHEP Study Group: Data Preservation in High Energy Physics**
DPHEP Study Group Collaboration (David M. South for the collaboration). 2013. 6 pp.
Published in PoS ICHEP2012 (2013) 536
Conference: C12-07-04 Proceedings
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Proceedings of Science Server; Link to Fulltext
Detailed record

3. **DPHEP: From Study Group to Collaboration**
DPHEP Collaboration (David M. South (DESY) for the collaboration). Sep 30, 2013. 6 pp.
Published in PoS DIS2013 (2013) 267
Conference: C13-07-18 Proceedings
e-Print: arXiv:1309.7868 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service; Proceedings of Science Server; Link to Fulltext
Detailed record

4. **Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics**
DPHEP Study Group Collaboration (Zaven Akopov (DESY) *et al.*). May 2012. 93 pp.
DPHEP-2012-001, FERMILAB-PUB-12-878-PPD
e-Print: arXiv:1205.4667 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server; ADS Abstract Service; OSTI Information Bridge Server; Fermilab Library Server (fulltext available); Link to Fulltext
Detailed record - Cited by 16 records

5. **Data Preservation in High Energy Physics**
DPHEP Study Group Collaboration (David M. South (DESY) for the collaboration). Jan 2011. 10 pp.
Published in J.Phys.Conf.Ser. 331 (2011) 012005
CHEP-2010
DOI: 10.1088/1742-6596/331/1/012005
Proceedings of plenary talk given at Conference: C10-10-18.4 Proceedings
e-Print: arXiv:1101.3186 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
ADS Abstract Service
Detailed record - Cited by 6 records

6. **Data Preservation in High Energy Physics**
DPHEP Study Group Collaboration (Richard Mount (SLAC) *et al.*). Nov 2009. 18 pp.
SLAC-R-987, DPHEP-2009-001, FERMILAB-PUB-09-856-CD
e-Print: arXiv:0912.0255 [hep-ex] | PDF
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
CERN Document Server ; ADS Abstract Service; SLAC Document Server; Fermilab Library Server (fulltext available); Link to Fulltext
Detailed record - Cited by 15 records

+ DPHEP@DESY documents

INSPIRE itself is a "level 1 data preservation project"

ZEUS
"Active Data Management Plan"

- wasn't called like that at the time, but
  a three page "bottom-up ZEUS ADMP"
  can be found in the
  2012 DPHEP study group document (see previous slide)

… and we conceptually implemented more or less exactly what we planned ☺ with some practical variations

(of course at that time it was already half way done)

"Discoverability"

DPHEP portal:

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

ZEUS web page:

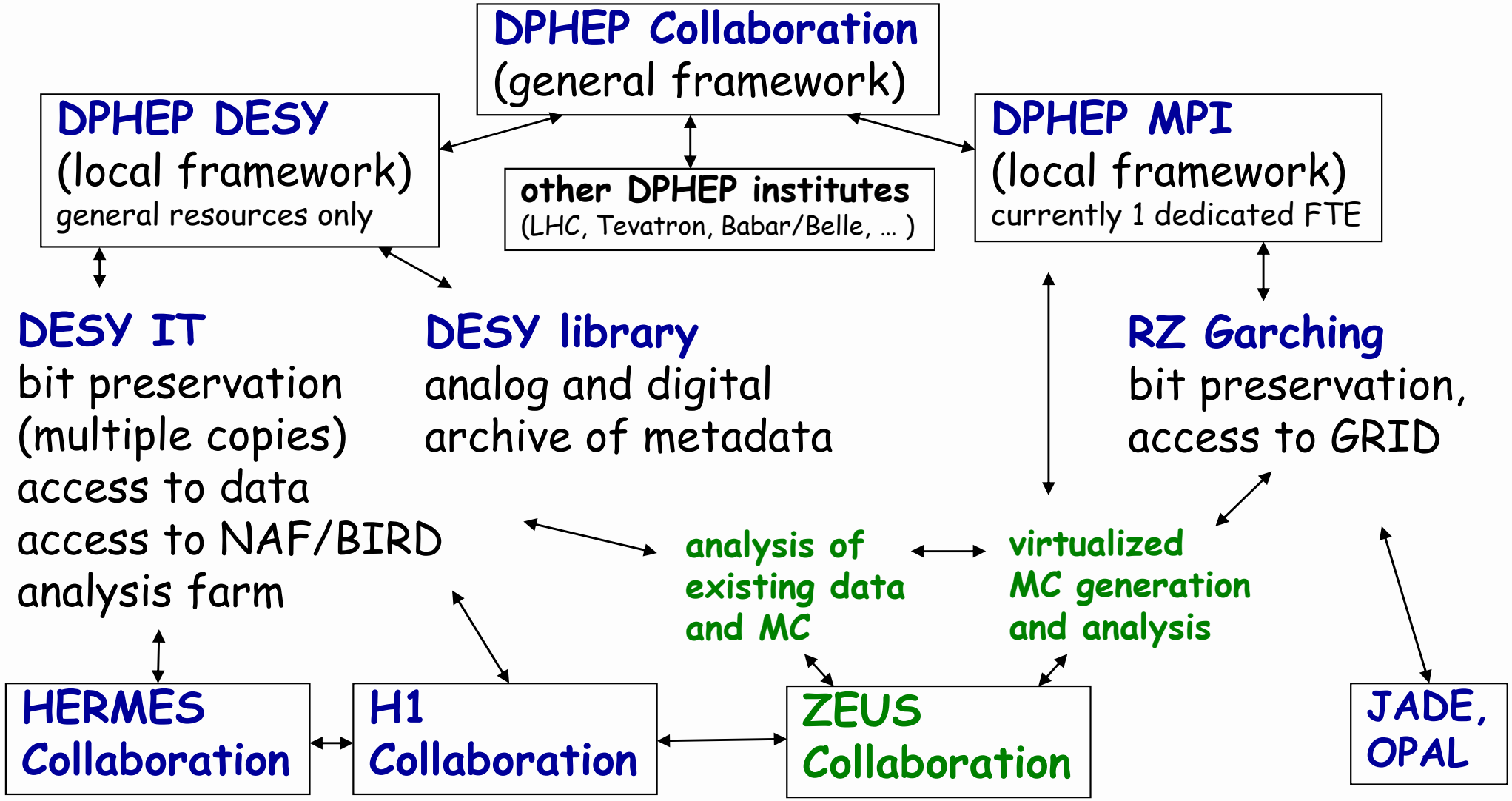
- http://www-zeus.desy.de/

information on ZEUS far from perfect

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

… but we are proud of what we achieved ☺

Challenge:
How to organize the Management?



Challenge: What is the "Data"?

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

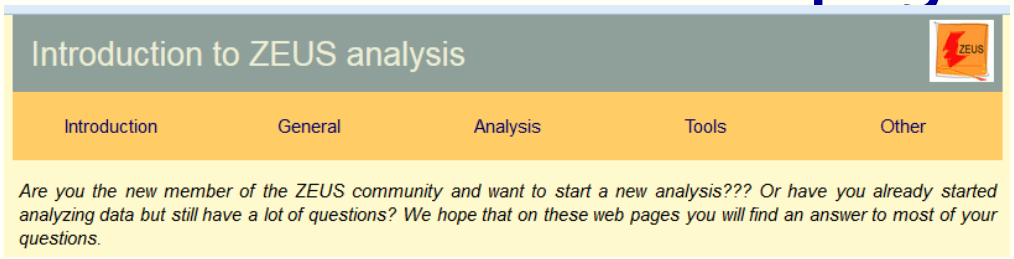
- see also https://wwwzeus.mpp.mpg.de

long term goal: ~0.5 FTE /year +IT

Analog and digital archive

- analog archive in DESY library

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



- knowledge preservation in "human neural networks" (ZEUS collaboration)

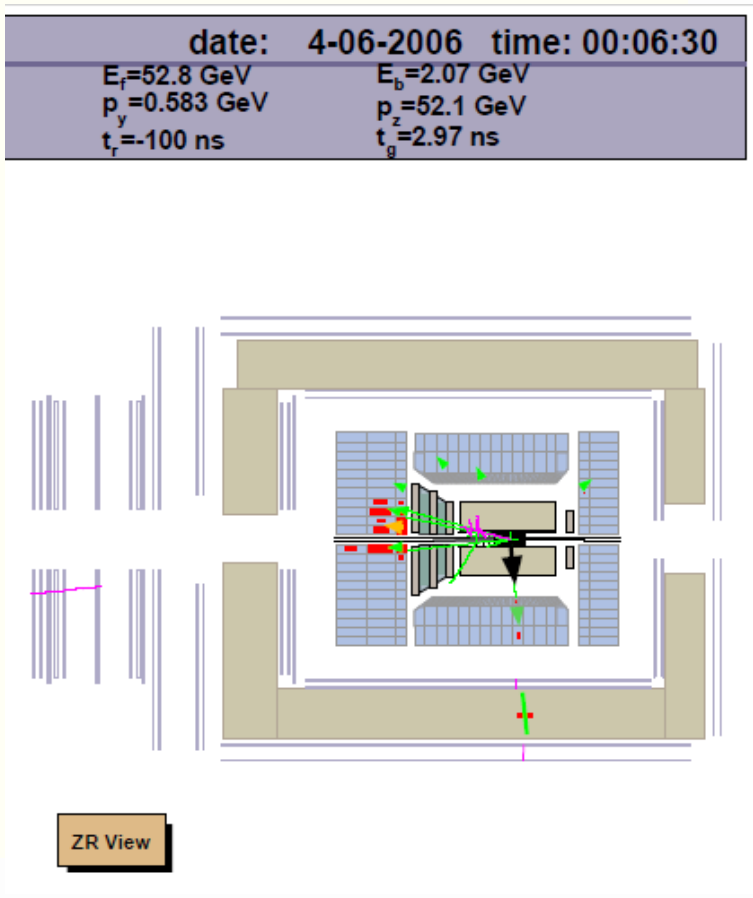
Common Ntuple analysis model

- **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, …)

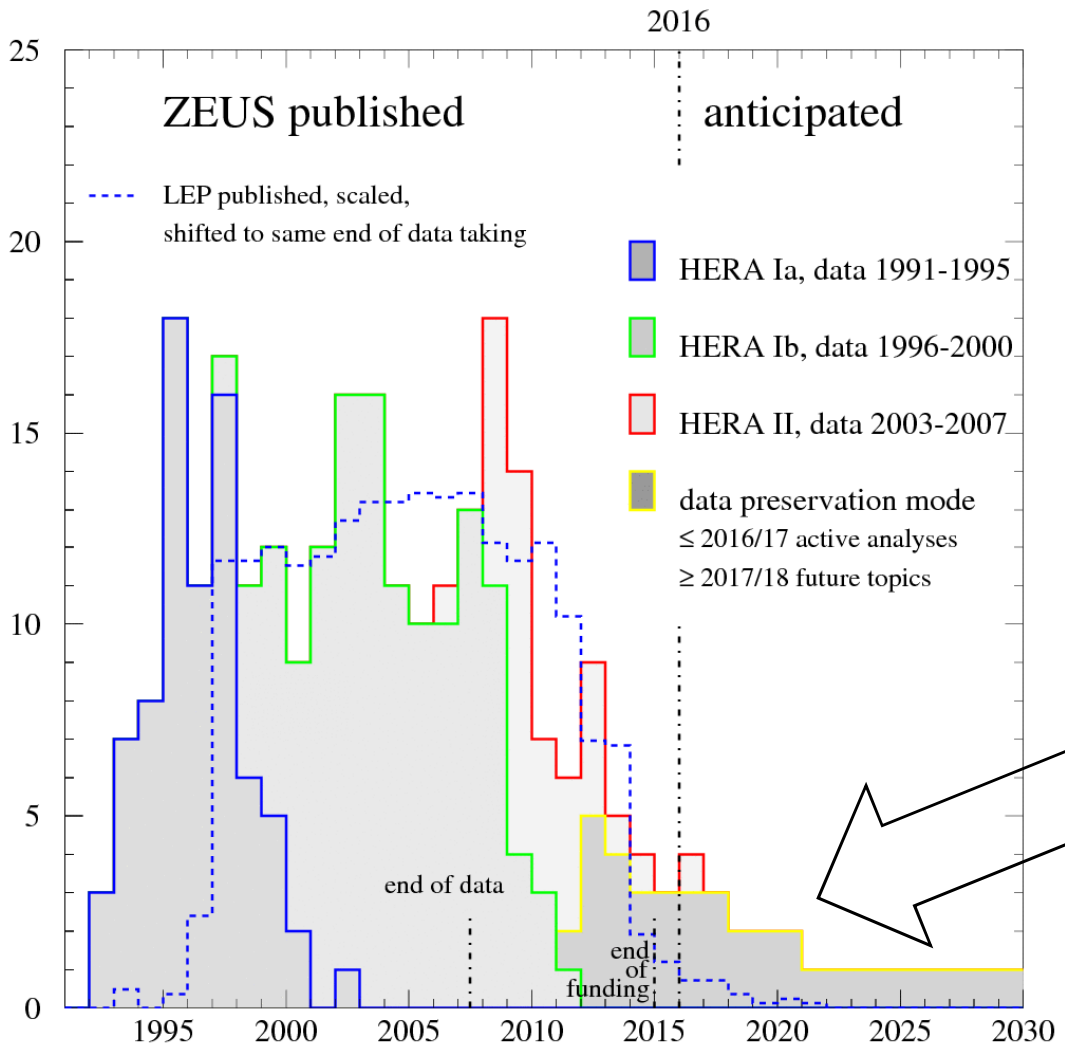
- **Well tested !**
  almost all recent ZEUS papers based on Common Ntuples

- **"Easy" to use**
  several recent ZEUS papers based on results produced by Master students from remote institutes, using resources at DESY

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



of ZEUS papers vs. time



scientific benefit of long term data preservation:

~10% of total benefit (for <<1% of total cost)

(my personal estimate - not official numbers)

difference between having/following a plan, or not having one

Challenge:
How to measure the success?

personal measure used on previous slide:

$$\frac{\text{expected \# of additional scientific papers}}{\text{total \# of scientific papers}}$$

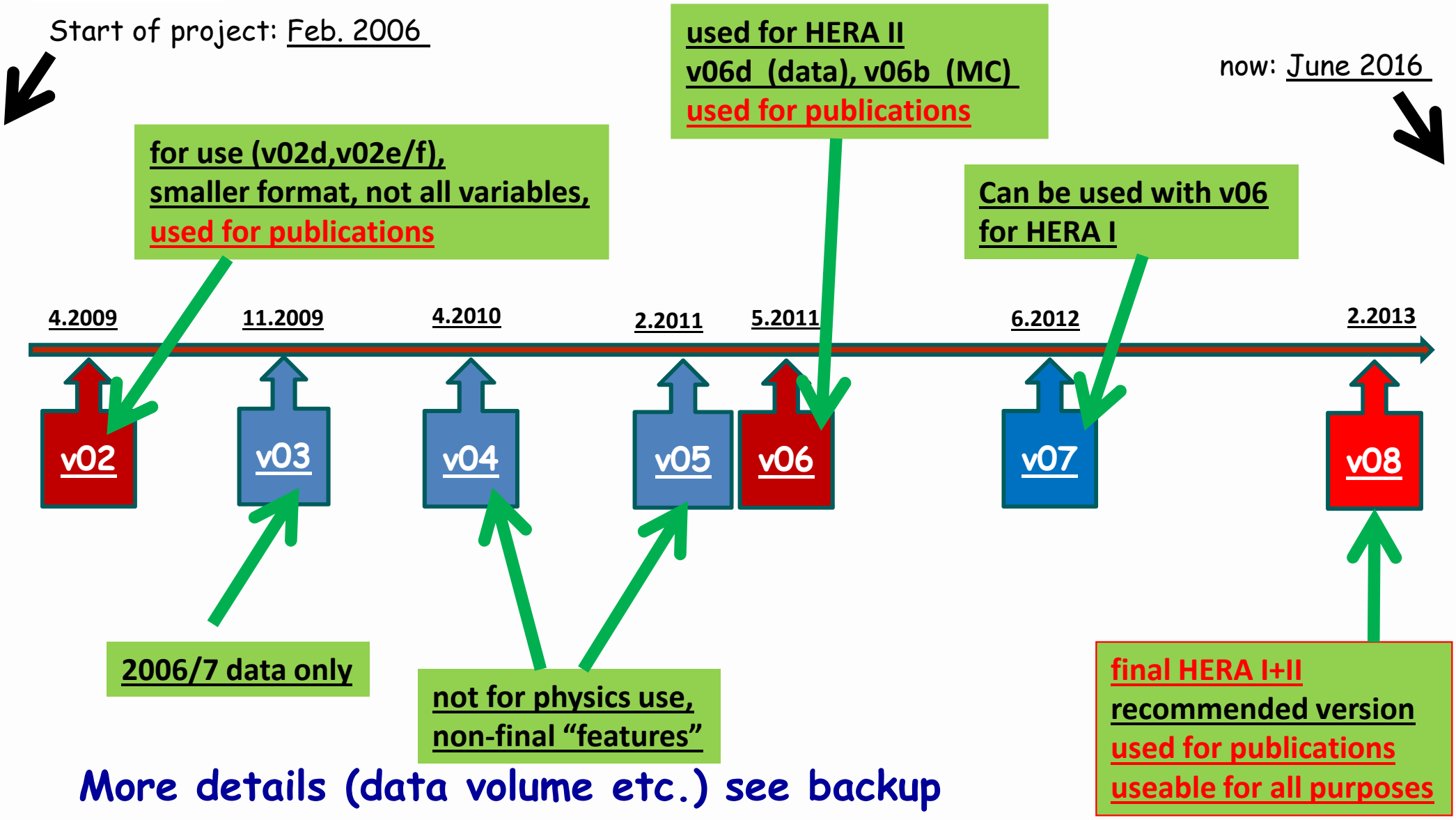
compared to

$$\frac{\text{estimated integrated cost of data preservation}}{\text{estimated integrated total cost of project}}$$

arguable - but is there a better one?

Available Common Ntuples

compiled by D. Szuba



More details (data volume etc.) see backup

Challenge:
"When will the project be finally done?"

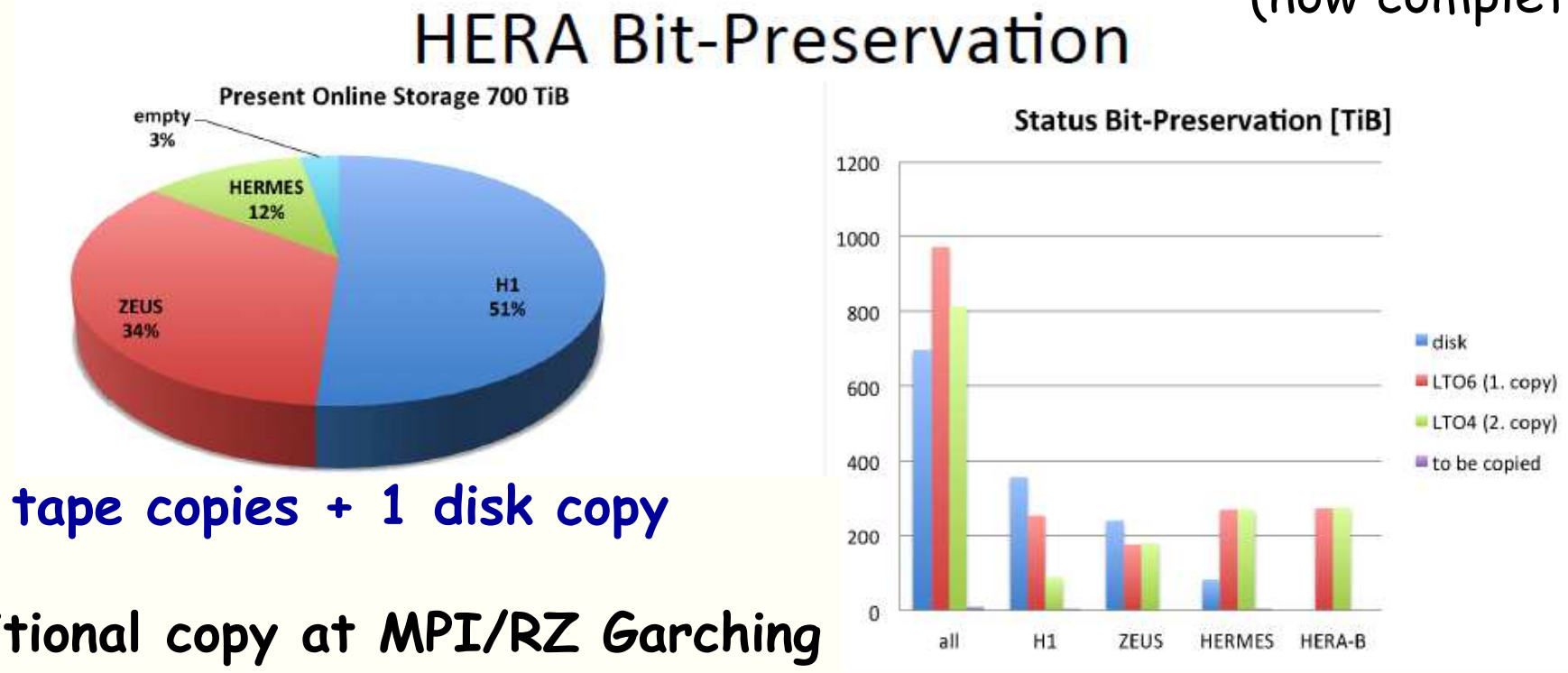
- my answer:
  (usually hard to digest for host labs, funding agencies, comittees …)

  if taken serious, a data preservation project will never be "done", unless and until one gives up on useability of the data

Challenge: Bit preservation

- at DESY: common approach for all three HERA experiments

status 06/2015
(now complete)



2 tape copies + 1 disk copy

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

Challenge: Virtualisation

- generation of new MC requires detector simulation
  - can not avoid use of full fledged "legacy" ZEUS software from the 1990's

- porting to new operating systems not realistic for ZEUS with realistic manpower

-> run existing software in "Virtual box" environment with "old" operating system

(implemented and maintained by MPI based on content development work at DESY)

- interface to new generators via HEPMC data format

Some ingredients for success of actual project

- Conceptually start the project about a decade before you need it ☺ ☺
  ZEUS: started 2006, full data preservation mode from Jan. 2015
- Boost your project by embedding it into a more global project ☺
  ZEUS: embedded into DPHEP, DESY DPHEP and MPI DPHEP
- Ensure strong support of collaboration/host lab during the implementation phase ☺ ☺
  ZEUS: strong support by collaboration management, strong initial support by DESY
  (~20 FTE-years ZEUS, + DESY IT/library)
- Make sure to have the necessary short term manpower and funding, as well as long term foresight ☺

Plans are long term, but people on short term contract/assignment often lack the perspective of/motivation to care about how things will be 5 years (or sometimes even only 1 month) after they leave or end their activity …

-> crucially need people with long term perspective to be involved

Some ingredients for success of actual project

- Make sure you start the 'user mode' well (>~ 2 years) before the temporary manpower ends (-> need to be able to fix "hickups" !) ☺

  ZEUS: user data preservation mode gradually started 2011-2013

- Ensure strong support of host lab or other funding body during the 'long term benefit' phase 😐

  ZEUS: scientific support OK, long term manpower/minimal funding support more difficult than expected/hoped for

- Make sure to get the necessary **dedicated long term manpower** (and funding!) going along with this support ☹

  ZEUS need: ~2/3 short term ~1/3 long term (~20 year integral)

  people understand the need to maintain storage, networks and tape vaults, and to provide some minimal CPU power, but rarely understand the (size of the) manpower need for **knowledge preservation, software preservation, and user support …**

  -> this is the main point upon which many of the past projects have failed and many of the current projects risk to fail

personal view

Comparison ZEUS data preservation/CMS open data

example for **synergy**!

- both use ROOT-based higher level data format ("level 3") with very similar content (CMS data format somewhat more complicated)
- both use virtual machine environments for software environment / preservation (well-developed for CMS open data analysis, under development by MPI for ZEUS for additional MC generation and analysis, not needed for baseline data and MC analysis at DESY)
- CMS more advanced in "transparent" access (open to general public)
- ZEUS more advanced in formal preservation of secondary information
- current practical synergy is that I work on both, can easily switch between the two (very similar conceptual approaches), and one profits from the other

"EU data principles"

- Discoverable: ZEUS and DPHEP web pages, conferences, workshops, …
- Accessible: ZEUS data are not (yet) open data

  (would need more manpower/funding)

  but "Free Access to ZEUS Data" programme for PhD students and physicists (e.g. EIC),
  data accessible at DESY, + on data grid via MPI
- Intelligible: bottleneck! currently OK, but would strongly profit from more manpower (keep experts involved!)
- Assessable: quality/reproduceability is ensured by the ZEUS collaboration
- Useable: Yes! (papers based on these data continously being published)
  all recent ZEUS papers are open access (DESY rule)

add:

- Sustainable: bottleneck! Can't do without some funding, in particular for long term manpower!
  "librarian" attitude to preservation could be useful!

Conclusions and Outlook

- HERA data are scientifically unique and worth preserving !

- 9 years after end of data taking in 2007, thanks to data preservation, ZEUS scientific output continues at a significant rate, for very little cost (expect ~10% of total scientific output, if long term sustainability is achieved). Made possible through substantial support by collaboration, host lab, and external institutes!

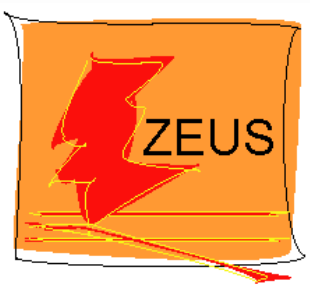
- ZEUS has successfully implemented its long term "Active Data Management Plan" for data preservation, worked out 2006-2012, and in full operation since Jan 2015. Integrated into DPHEP strategy.

- Bottleneck: Long term data preservation needs long term manpower: don't need "much" (~50% addition to investment for data preservation implementation, ~O(‰) of original project investment, spread over 20 years), but 0 will not do …

Backup

Size of data sets

compiled by D. Zotkin/A.G.



Root files (officially preserved) units: Tb (status 4.9.13)

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 flavours, vector mesons, …)

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

~ 4 billion MC events, for almost any analysis

generation of additional MC samples might be possible (see talk A. Verbytskyi)

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

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

ZEUS data preservation status

All relevant ZEUS data + MC on DESY online store + two tape copies

User analysis on NAF/BIRD ongoing well

ZEUS archive web server now the default

Additional data copy + analysis setup at MPI/RZ Garching,
including possibility to simulate new MC