Data Preservation at DESY.

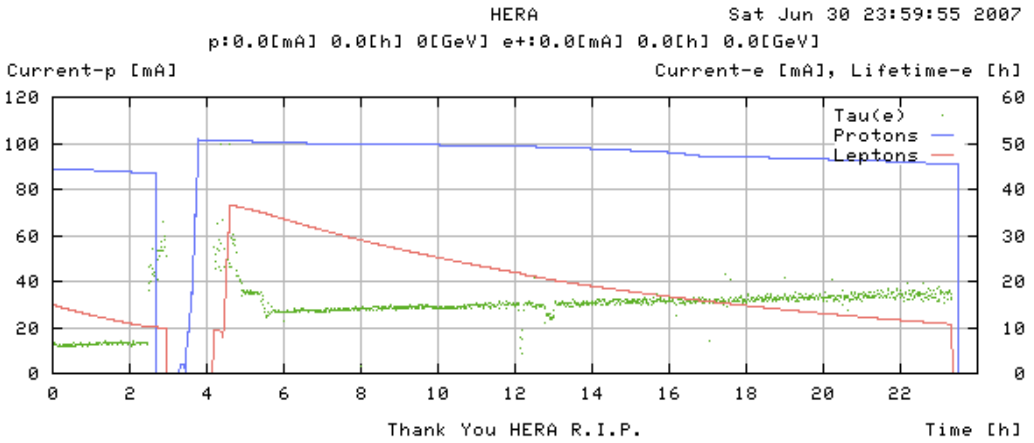
Update from the DESY-DPHEP Group

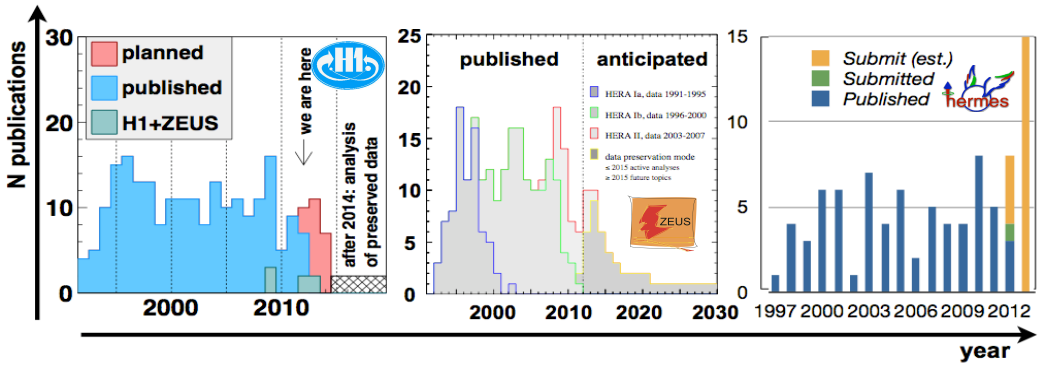
David South (DESY)

FNAL-DESY-SLAC DPHEP Workshop
Fermilab, 28th March 2014

The end of data taking at HERA: June 30th 2007

- Unique period of time in HEP history: change from many running experiments of various types to essentially only one
- HERA, stopped taking data 6.5 years ago – so what's happened since then?
- Much like LEP before us and seen by BaBar, publications still continue well after data taking: *~25% of total so far!*
  - H1: 55 papers since June 2007, out of a total of 218
  - ZEUS: 64 out of a total of 241





DPHEP activity at DESY since 2008

- The first few years after data taking: 2008-2010
  - Formation of initial ideas, first DPHEP workshops
  - Grand surveys done: data, hardware, software, technologies
  - Establishing the physics case for data preservation
  - Defining the DHPEP preservation levels: HERA experiments plan for level 3-4
  - Finding the people to do the work

DPHEP activity at DESY since 2008

- The first few years after data taking: 2008-2010
  - Formation of initial ideas, first DPHEP workshops
  - Grand surveys done: data, hardware, software, technologies
  - Establishing the physics case for data preservation
  - Defining the DHPEP preservation levels: HERA experiments plan for level 3-4
  - Finding the people to do the work
- Key areas of activity at DESY since 2011
  1. Preparation of the data for preservation and archival storage of the data themselves
  2. Data preservation: really preservation of software + environment: the `sp-system`
  3. Documentation: INSPIRE, digital meta-data and non-digital material
  4. Governance, future collaboration structures and open access/public data, outreach

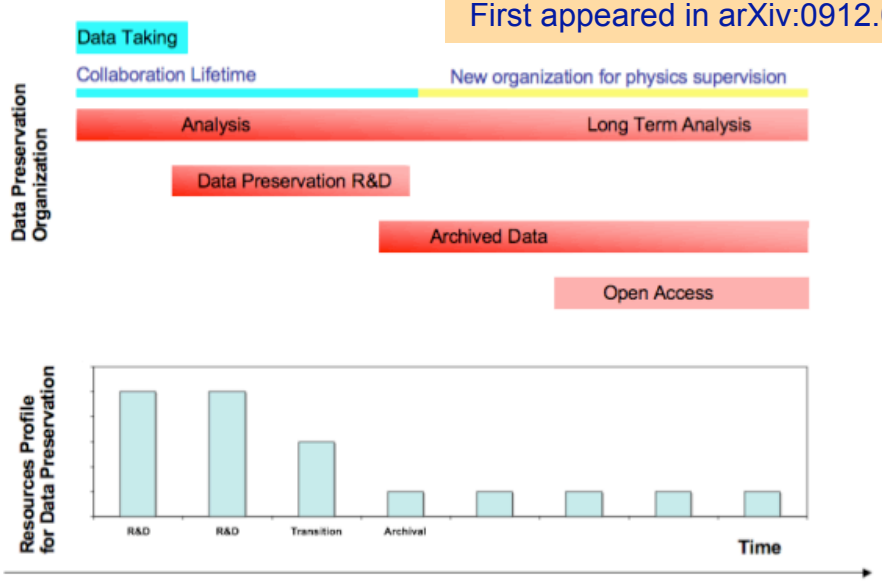
DPHEP activity at DESY since 2008

- The first few years after data taking: 2008-2010
  - Formation of initial ideas, first DPHEP workshops
  - Grand surveys done: data, hardware, software, technologies
  - Establishing the physics case for data preservation
  - Defining the DHPEP preservation levels: HERA experiments plan for level 3-4
  - Finding the people to do the work
- Key areas of activity at DESY since 2011
  1. Preparation of the data for preservation and archival storage of the data themselves
  2. Data preservation: really preservation of software + environment: the `sp-system`
  3. Documentation: INSPIRE, digital meta-data and non-digital material
  4. Governance, future collaboration structures and open access/public data, outreach
- Some things have been completed, typically well defined tasks such as the documentation (rather specialised person-power), other things still on going, including final preparation of archival data storage

The DESY-DPHEP Group

First appeared in arXiv:0912.0255

- During first years, regularly more than a dozen people involved
- Group made up of people from H1, ZEUS and HERMES as well as DESY-IT and DESY-Library
  - The available person-power has declined, in line with the model presented in the first DPHEP publication
  - 2014: There are now only a couple of people involved at DESY



- Initial person-power estimates included provision for support in 2014 and beyond

		2011	2012	2013	Translates into Position	2014++
DESY-IT	Validation	1.0		0.5	3 year FTE 2011 – 2013	(0.5)
	Storage		1.0	0.5		
H1	Validation	0.5	1.0	0.5	2 year extention for 2011 – 2013	(0.5)
	Documentation	0.5	0.5		1 year extention for 2012	
ZEUS	Validation	0.5	1.0	0.5	(Initial) 2 year FTE 2011 – 2013	(0.5)
	Documentation	0.5	0.5		1 year FTE 2011 – 2012	
HERMES	Validation			0.5	0.5 year extention for 2013	(0.5)
	Documentation	0.5	0.5		1 year FTE 2011 – 2012	

- Long term support has proven difficult to secure, especially when trying to find the right people for the job
- All current DP person-power for the experiments runs out **this year**

Key area 1: Data for preservation and archival storage

- Deciding which data (and MC) are needed for the long term depends on the preservation model assumed: Level 4 goes back to the raw data
- Final production of HERA data for preservation only completed last year; majority of MC production expected to be concluded this year
- Estimates for final **DPHEP dataset** volume ready (including MC samples)
  - Plan calls for two tape copies and an "always online" (disk) component
  - Data which should be archived, but not online all the time: re-pack into larger files
  - Costs not prohibitive on data volume basis

Expt	Online (TB)	Total (TB)
H1	250	500
ZEUS	250	250
HERMES	100	300
Total	**600**	**1050**
HERA-B	?	300

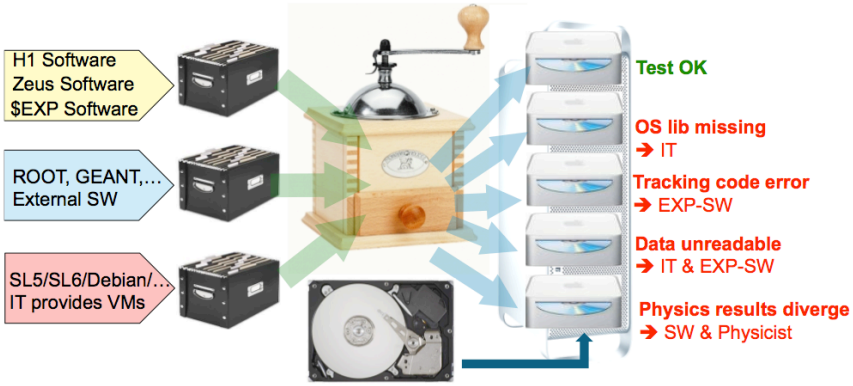
Different strategies visible

- Dedicated system too costly in both hardware and support required
  - All collaborations use dCache for mass storage and this system will continue at DESY-IT for the LHC, photon-physics and others. Natural solution for DPHEP dataset
  - Changes "transparent" for user, relying on work by DESY-IT

Key area 2: Software preservation & validation: `sp-system`

Write up: arXiv:1310.7814

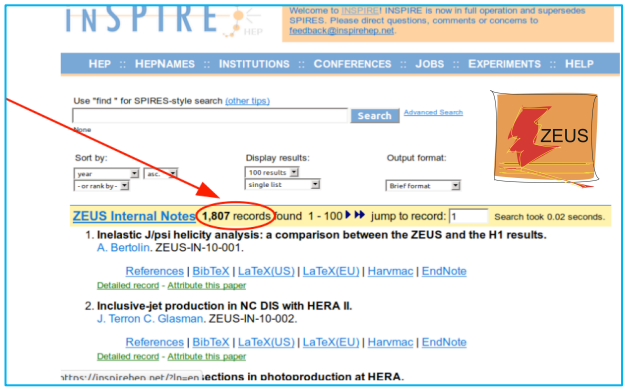
- Fairly early on, HERA experiments decided to try to migrate software for as long as possible rather than freezing the current environment
- Pilot project of a system for software preservation and validation in 2010

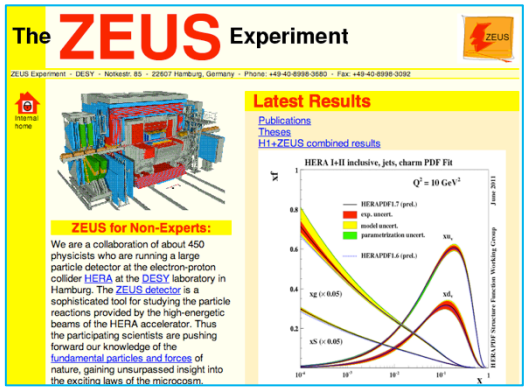


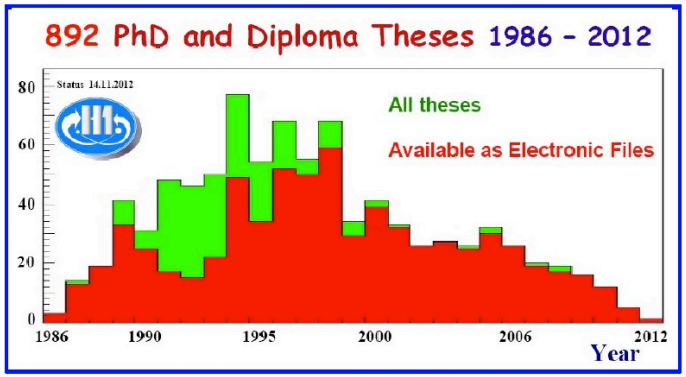
- *Briefly*: The idea of the `sp-system` is to help perform migrations to newer software versions and environments, where transitions are performed often and validated by a comprehensive set of tests provided by the expts
  - The output of such a system is a **recipe** for deployment on (future) external resource(s)
  - Future analysis resources maybe local batch farm, grid, cloud, whatever
  - *The idea is **not** to run analysis within the system itself!*
- Due to available resources and changes in personnel, implementation at DESY is still not in production mode
  - Project is rather ambitious and has taken longer than anticipated: definition of tests essentially done, but still requires much work to be done on the validation side

Key area 3: Documentation

- Successful collaboration between **INSPIRE**, the experiments and the DESY Library
- **Digital documentation** such as web-pages revised, reduced and streamlined for future use
- Lots of effort done sorting the vast amount of **non-digital documentation**
- Work done by key people with the *right expertise* and *experience* for the job

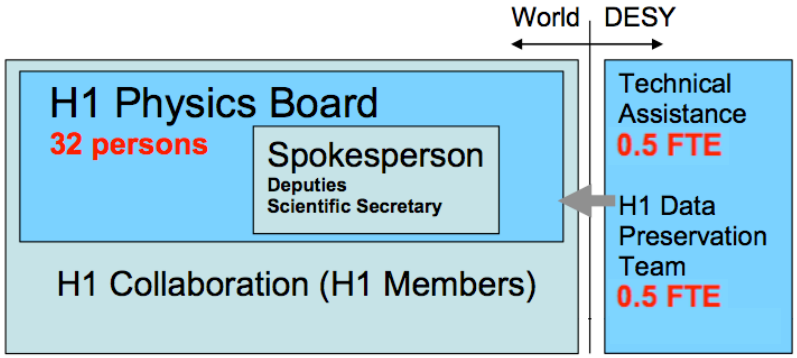




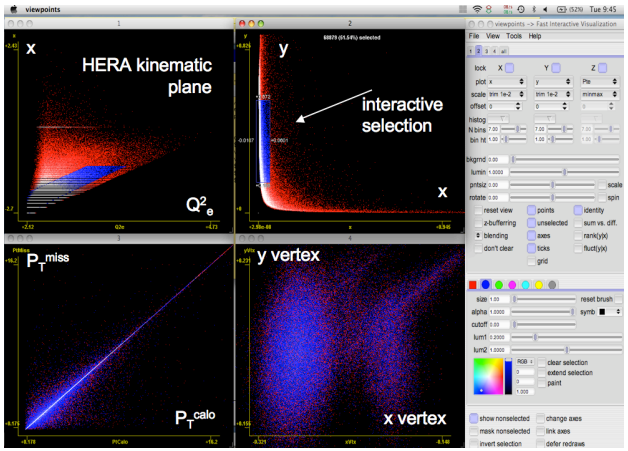


Key area 4: Governance, open access and outreach

- H1 collaboration moved to a **new management model** in July 2012
  - Formation of ***H1 Physics Board***, to replace Collaboration Board (institute based)
  - Future author list policies also set down in new constitution approved by collaboration



- ZEUS and HERMES management teams retain same model as before, but similarly to H1 the collaborating institute layer is now removed
  - Remaining physics ZEUS working groups consolidated to a single physics group
- **Open access** still to be considered and/or defined by the HERA experiments
- **Outreach** is a great idea, but was not possible without dedicated resources
  - Already dropped in 2011 table shown earlier
  - Ideas existed, but nothing concrete came of it



In summary: how are we doing?

- Key Area 1: Data for preservation and archival storage
  - Activity on-going, DPHEP data expected to be on new, long term data storage by the summer, and in use by the end of the year
- Key Area 2: Software preservation & validation: `sp-system`
  - Still much to be done, will be difficult in 2015 without dedicated person-power in the experiments. However, a **new position** will begin in DESY-IT in May for an (initial) 2 years
- Key Area 3: Documentation
  - This is pretty much concluded now, final discussions on future model of webpages
- Key Area 4: Governance, open access and outreach
  - Here not much concrete has happened at all..
- A lesson from HERMES: Financial support officially ended December 31st, 2012
  - In 2012 they tried to finish off as much as possible, physics results and data preservation
  - Hardware turn-off and transfer to DESY-IT central services completed; Validation project within `sp-system` not really implemented
  - Current situation: no dedicated manpower for any HERMES activities; the same will apply to H1 and ZEUS, at least for data preservation, at the end of 2014

DESY

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!

- There is a **great reduction** in person power (and available expert knowledge) as well as funding as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!

- There is a **great reduction** in person power (and available expert knowledge) as well as funding as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!

- There is a **great reduction** in person power (and available expert knowledge) as well as funding as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends
  - **Don't underestimate the required person-power**: for funding or practical reasons

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!

- There is a **great reduction** in person power (and available expert knowledge) as well as funding as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends
  - **Don't underestimate the required person-power**: for funding or practical reasons
  - **Dedicated manpower** is needed, people working on this part time or in spare time is not enough: such initiatives cannot "run for free"

Conclusions from DPHEP 8 Workshop at CERN in January: Some lessons learned from Data Preservation @ DESY

- The **physics output** tail seen by LEP also rings true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**

- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1-HERA1 took 3 years to get 3 months of work)

- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!

- There is a **great reduction** in person power (and available expert knowledge) as well as funding as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends
  - **Don't underestimate the required person-power**: for funding or practical reasons
  - **Dedicated manpower** is needed, people working on this part time or in spare time is not enough: such initiatives cannot "run for free"
  - Losing the best people for the best roles is almost inevitable and finding support for unfinished things is extremely difficult. Difficult to capture the best candidates without providing a **long term perspective**, the need for which despite established physics cases, has been **difficult to convince people of**

How could DESY, FNAL and SLAC cooperate (more)?

- DPHEP is now run from CERN and the focus is on the LHC experiments now, who have (more) time
  - It seems to be the majority of any funding secured will also flow in this direction
  - How can DESY, FNAL and SLAC still contribute?
  - We were of course the main contributors to the 2012 paper..
  - Do we want to write up lessons learned / experience gained in HEP data preservation?
- The DPHEP Collaboration agreement will help us stay involved
  - DESY is now in the final part of the signing procedure, this will happen very soon
  - Can we at least update contact names and/or propose signatories for FNAL and SLAC?
- Then what about *real collaboration*?
  - Do other experiments want to get involved in validation systems? (LHC also looking..)
  - And what else might be possible?