



DPHEP Summary

Data Sharing – In Time and Space

Jamie.Shiers@cern.ch

School on Grid & Cloud Computing



International Collaboration for Data Preservation and
Long Term Analysis in High Energy Physics

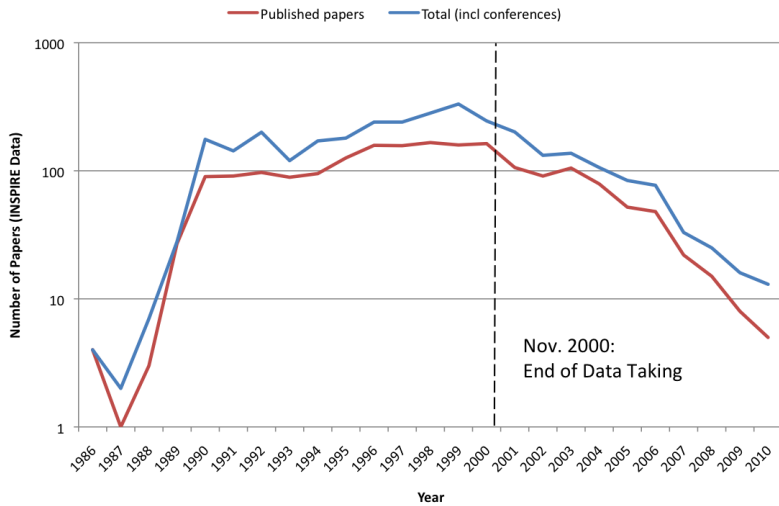
Training Overview

1. Background to Long-Term Data Preservation in HEP – the DPHEP Study Group
2. “2020 vision” for DP in HEP
- 3. DP in other disciplines – how we can benefit significantly from work (models, standards, procedures, wisdom, tools, services etc.) of others – the bulk of the material comes from other projects / disciplines**
- 4. A strategy for DP in HEP**

2020 Vision for LT DP in HEP

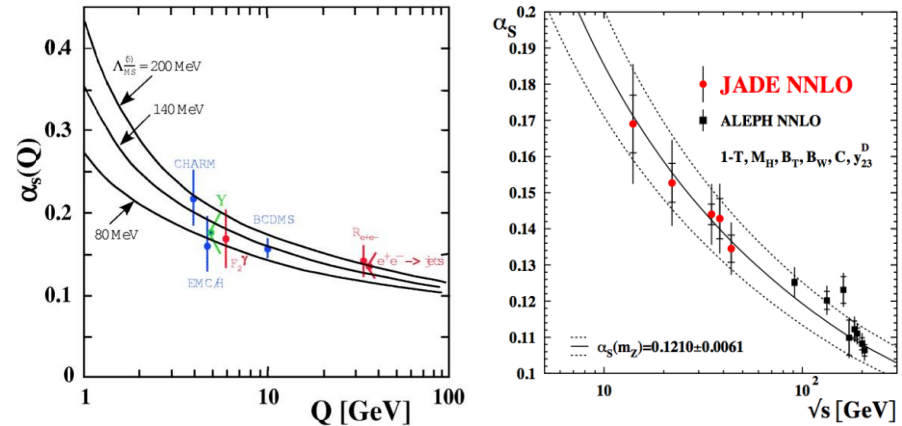
- Long-term – e.g. FCC timescales: **disruptive change**
 - By **2020**, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to **annotate** further
 - Best practices, tools and services well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards
 - **DPHEP portal**, through which data / tools accessed
- **Agree with Funding Agencies clear targets & metrics**

1 - Long Tail of Papers



3

2 - New Theoretical Insights

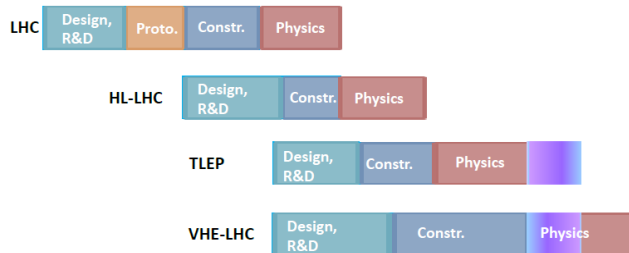


4

3 - "Discovery" to "Precision"



possible long-term time line



5

Use Case Summary

1. Keep data usable for ~1 decade
2. Keep data usable for ~2 decades
3. Keep data usable for ~3 decades

Volume: 100PB + ~50PB/year (+400PB/year from 2020)

7

Collaboration – Benefits

- In terms of 2020 vision, collaboration with other projects has arguably advanced us (in terms of implementation of the vision) by several years
- **I typically quote 3-5 years and don't think that I am exaggerating**
- **Concrete examples include “Full Costs of Curation”, as well as proposed “Data Seal of Approval+”**
- With or without project funding, we should continue – and even strengthen – this collaboration
 - APA events, iDCC, iPRES etc. + joint workshops around RDA
- **The HEP “gene pool” is closed and actually quite small – we tend to recycle the same ideas and “new ones” sometimes needed**

APARSEN Training & Knowledge Base

Home > About APARSEN > APARSEN Webinars

APARSEN holds regular webinars on emerging topics/results in the project. These webinars all have a similar setting and format and take place as virtual meetings with APARSEN partners and external stakeholders participating. Their aim is to present some of the results of APARSEN when a topic milestone is reached and to foster discussions with stakeholders outside of APARSEN.

Participation is free, no registration required!

APARSEN Webinars

- Storage Solutions for Digital Preservation
- Long Term Preservation and Digital Rights Management
- Certification of Digital Preservation Repositories**
- Interoperability of Persistent

Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the Office of Science for research funding are required to include a Data Management Plan (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be **shared and preserved** or explains why data sharing and/or preservation are not possible or scientifically appropriate.
- At a minimum, DMPs must describe how data sharing and preservation will enable **validation of results**, or how results could be validated if data are not shared or preserved.
- Similar requirements from European FAs and EU (H2020)

How to respond?

- a) Each project / experiment responds to individual FA policies
 - $n \times m$

- b) We agree together – service providers, experiments, funding agencies – on a common approach
 - DPHEP can (should?) help coordinate

- b) almost certainly (much) cheaper / more efficient but what does it mean in detail?

- Open Archival Information System reference model provides:
 - fundamental concepts for preservation
 - fundamental definitions so people can speak without confusion
 - *“now adopted as the de facto standard for building digital archives”*
 - In *Cyberinfrastructure Vision for 21st Century Discovery*
 - ▶ <http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf>





Data Seal of Approval: Guidelines 2014-2015

Guidelines Relating to Data Producers:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.
- 2. The data producer provides the data in formats recommended by the data repository.**
3. The data producer provides the data together with the metadata requested by the data repository.

1. DPHEP Portal

2. **Digital library** tools (**Invenio**) & services (**CDS, INSPIRE, ZENODO**) + domain tools (**HepData, RIVET, RECAST...**)
3. **Sustainable software**, coupled with advanced **virtualization** techniques, “snap-shotting” and **validation** frameworks
4. **Proven bit preservation** at the 100PB scale, together with a **sustainable** funding model with an outlook to 2040/50
(and several EB of data)
5. **Open Data**

DPHEP Portal – Zenodo like?

The screenshot shows the Zenodo website interface. At the top, there is a search bar with a magnifying glass icon and a 'Search' button. Below the search bar, there is a 'Filter by types' button. The main content area is titled 'Recent Uploads' and displays a list of items. Each item includes a date, a category (e.g., 'Journal article', 'Thesis'), and a 'View' button. The first item is 'The e-book phenomenon: a disruptive technology' by Tom D. Wilson, dated 01 April 2014. The second item is 'Χρόνιες Επιδράσεις του Καπνίσματος στη Λειτουργική Ικανότητα του Κυκλοφορικού Συστήματος Νεαρών Υγιών Ατόμων' by Papathansiou, George ; Evangelou, Angelos, dated 10 November 2010. The third item is 'Archaeobotanical remains from Mitchelstown and Ballnamona', dated 30 March 2014. On the right side, there is a 'GitHub integration' section with a GitHub logo and a 'New to ZENODO?' section with a list of features.

zenodo.org

Search

Filter by types

Recent Uploads

01 April 2014 **Journal article** **Open access** [View](#)

The e-book phenomenon: a disruptive technology
Tom D. Wilson

The emergence of the e-book as a major phenomenon in the publishing industry is of interest, world-wide. The English language market, with Amazon.com as the major player in the market may have dominated attention, but the e-book has implications for many ...

10 November 2010 **Thesis** **Open access** [View](#)

Χρόνιες Επιδράσεις του Καπνίσματος στη Λειτουργική Ικανότητα του Κυκλοφορικού Συστήματος Νεαρών Υγιών Ατόμων
Papathansiou, George ; Evangelou, Angelos

Εισαγωγή Το κάπνισμα αποτελεί τον σοβαρότερο (ίσως παράγοντα κινδύνου μελλοντικής καρδιοαγγειακής νοσηρότητας και θνητότητας ενώ θεωρείται ως η κυριότερη αντιστρεπτή αιτία θανάτου. Το κάπνισμα συνδέεται με χρονότροπη καθυστέρηση λόγω δυσλειτουργίας του ...

Uploaded by [George](#) on 30 March 2014.

30 March 2014 **Report** **Open access** [View](#)

Archaeobotanical remains from Mitchelstown and Ballnamona

GitHub integration

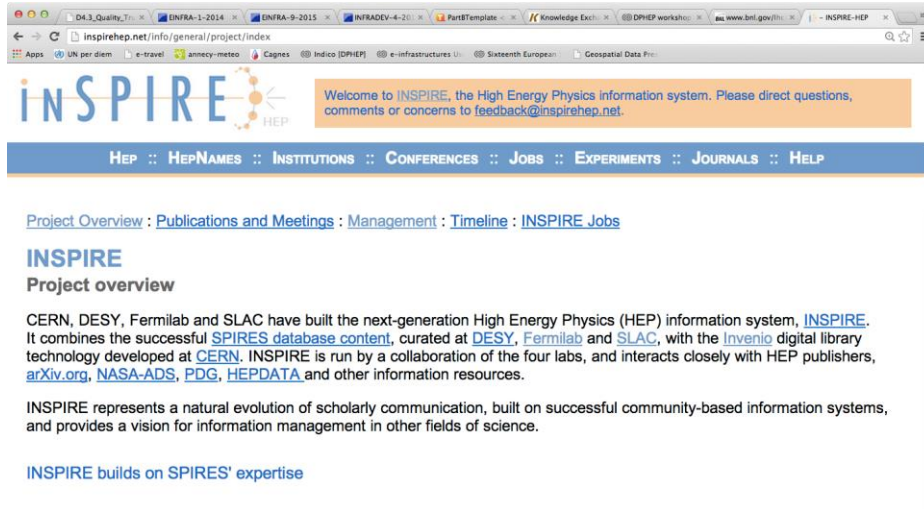
Want to preview the public beta of GitHub integration? Just [Sign In](#) with your GitHub account and [click here](#).

New to ZENODO?

- **Research. Shared.** – all research outputs from across all fields of science are welcome!
- **Citeable. Discoverable.** – uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- **Community Collections** – accept or reject uploads to your own community collections (e.g workshops, EU projects or your complete own digital repository).
- **Funding** – integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- **Flexible licensing** – because not everything is under Creative Commons.
- **Safe** – your research output is stored safely for the future in same cloud infrastructure as research data from CERN's Large Hadron Collider.
- **DropBox integration** – upload files straight from your DropBox.

Documentation projects with INSPIREHEP.net

- Internal notes from all HERA experiments now available on INSPIRE
 - A collaborative effort to provide “consistent” documentation across all HEP experiments – starting with those at CERN – **as from 2015**
 - (Often done in an inconsistent and/or ad-hoc way, particularly for older experiments)



Welcome to INSPIRE, the High Energy Physics Information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

[Project Overview](#) : [Publications and Meetings](#) : [Management](#) : [Timeline](#) : [INSPIRE Jobs](#)

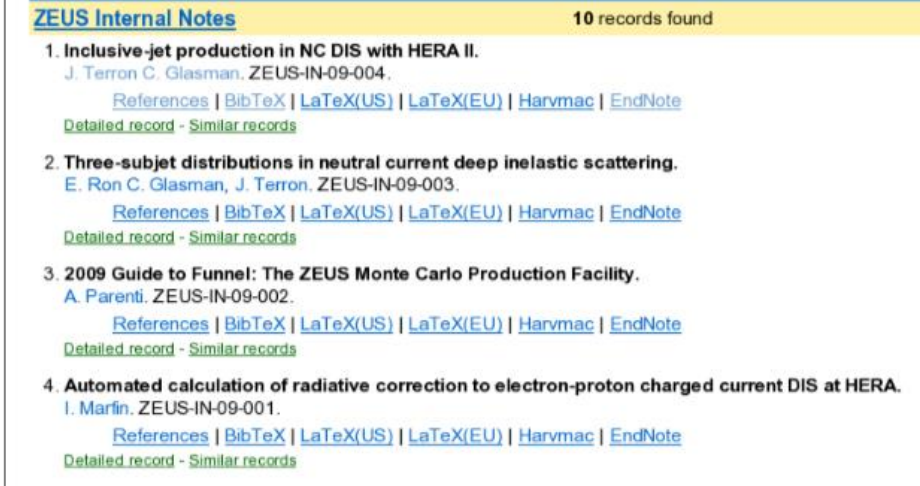
INSPIRE

Project overview

CERN, DESY, Fermilab and SLAC have built the next-generation High Energy Physics (HEP) information system, [INSPIRE](#). It combines the successful [SPIRES database content](#), curated at [DESY](#), [Fermilab](#) and [SLAC](#), with the [Invenio](#) digital library technology developed at [CERN](#). INSPIRE is run by a collaboration of the four labs, and interacts closely with HEP publishers, [arXiv.org](#), [NASA-ADS](#), [PDG](#), [HEpdata](#) and other information resources.

INSPIRE represents a natural evolution of scholarly communication, built on successful community-based information systems, and provides a vision for information management in other fields of science.

INSPIRE builds on SPIRES' expertise



ZEUS Internal Notes 10 records found

- Inclusive-jet production in NC DIS with HERA II.**
J. Terron C. Glasman. ZEUS-IN-09-004.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)
- Three-subjet distributions in neutral current deep inelastic scattering.**
E. Ron C. Glasman, J. Terron. ZEUS-IN-09-003.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)
- 2009 Guide to Funnel: The ZEUS Monte Carlo Production Facility.**
A. Parenti. ZEUS-IN-09-002.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)
- Automated calculation of radiative correction to electron-proton charged current DIS at HERA.**
I. Marfin. ZEUS-IN-09-001.
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)
[Detailed record](#) - [Similar records](#)



Summary

- It would be misleading to present DP in HEP as a “solved problem” – it is not
- However, many of the building blocks are understood with corresponding services, tools and support units
- A strategy, building on certified repositories and generic tools, complemented by additional metrics, is being elaborated
- Its still only 2014 – good progress expected in coming 2-3 years – well ahead of “2020”!