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

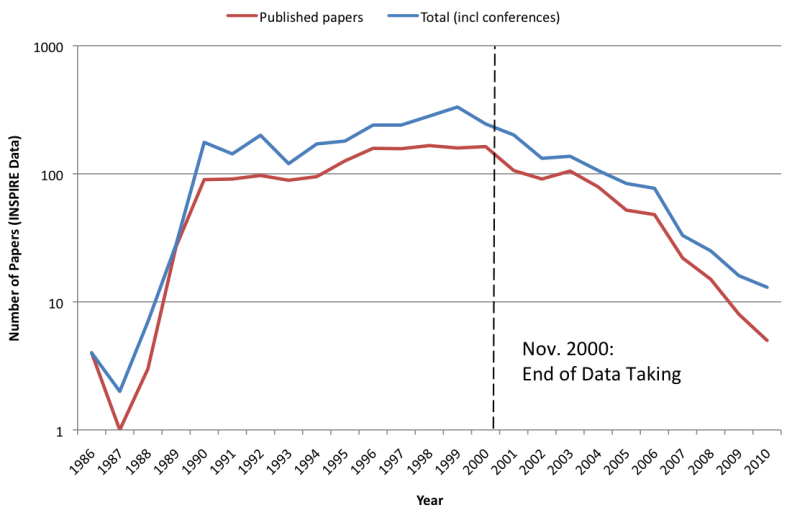
1. Background to Long-Term Data Preservation in HEP – the DPHEP Study Group
2. "2020 vision" for DP in HEP
3. **DP in other disciplines – how we can benefit significantly from work (models, standards, procedures, wisdom, tools, services etc.) of others – the bulk of the material comes from other projects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

2020 Vision for LT DP in HEP

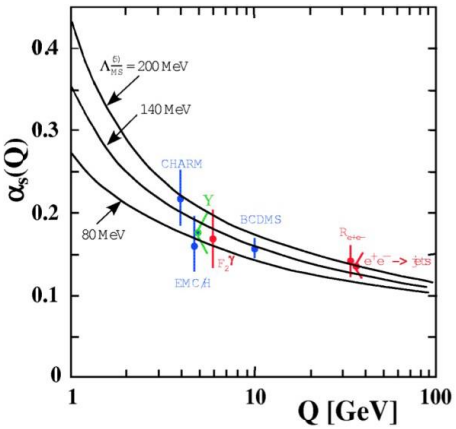
- ***Long-term – e.g. FCC timescales*: *disruptive change***
  - By **2020**, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to **annotate** further
  - 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

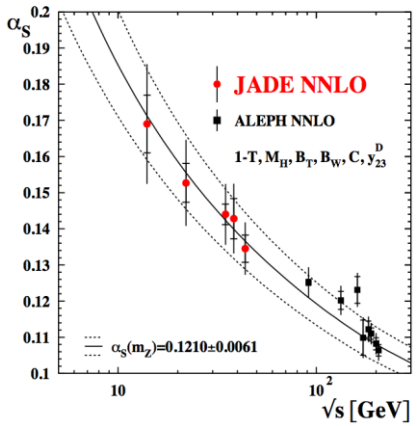
➢ **Agree with Funding Agencies clear targets & metrics**

1 – Long Tail of Papers



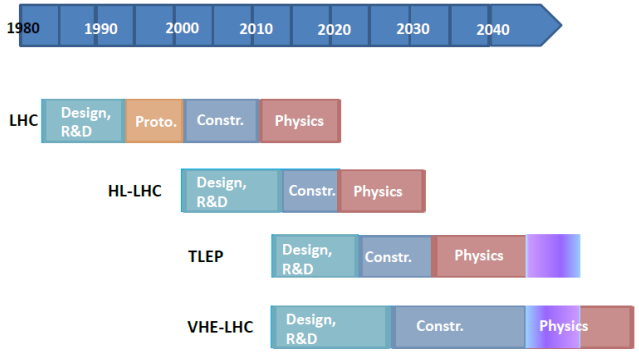
2 – New Theoretical Insights





3 – "Discovery" to "Precision"

possible long-term time line



Zimmermann

Use Case Summary

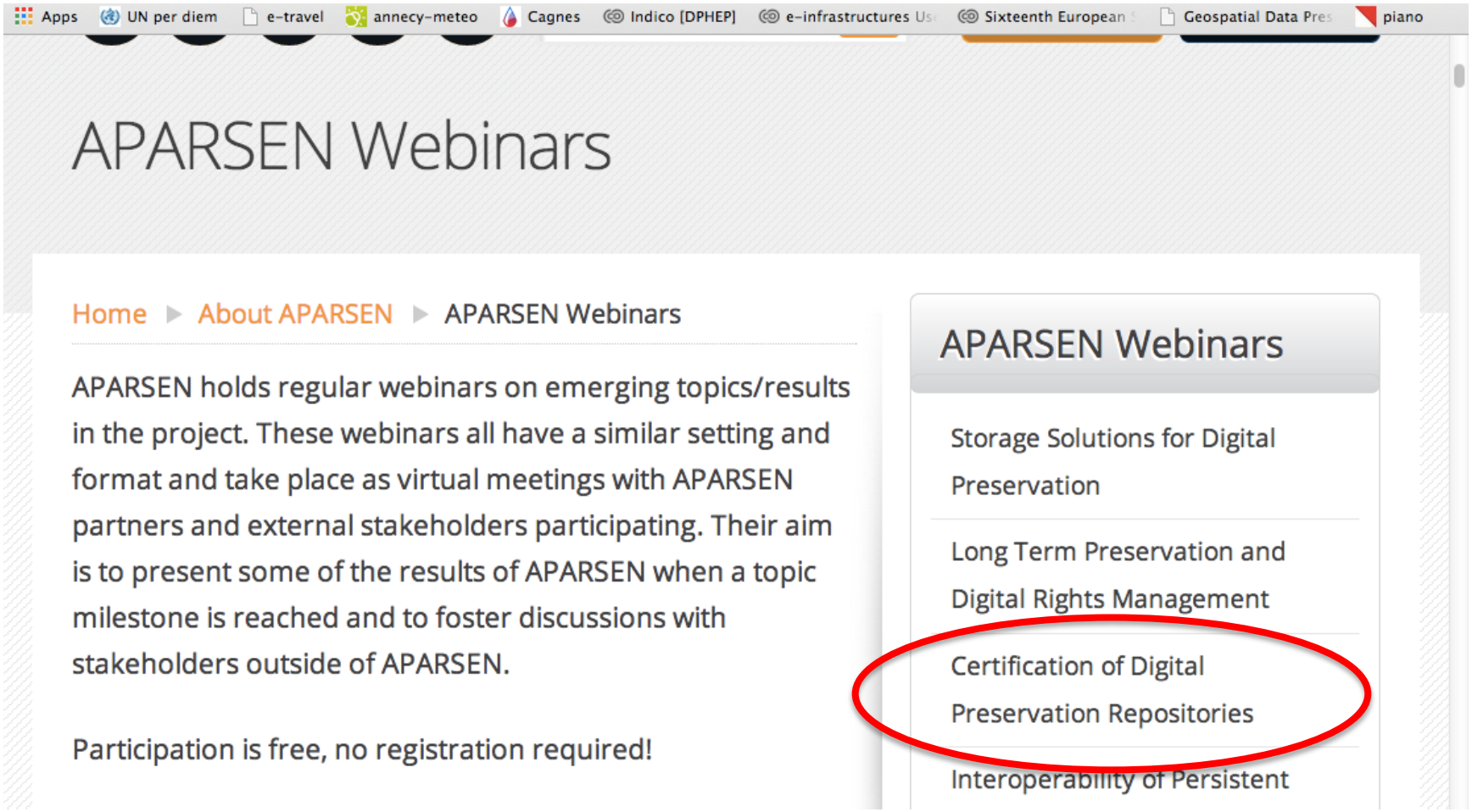
1. Keep data usable for ~1 decade
2. Keep data usable for ~2 decades
3. Keep data usable for ~3 decades

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

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



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 x 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?**

OAIS

- 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

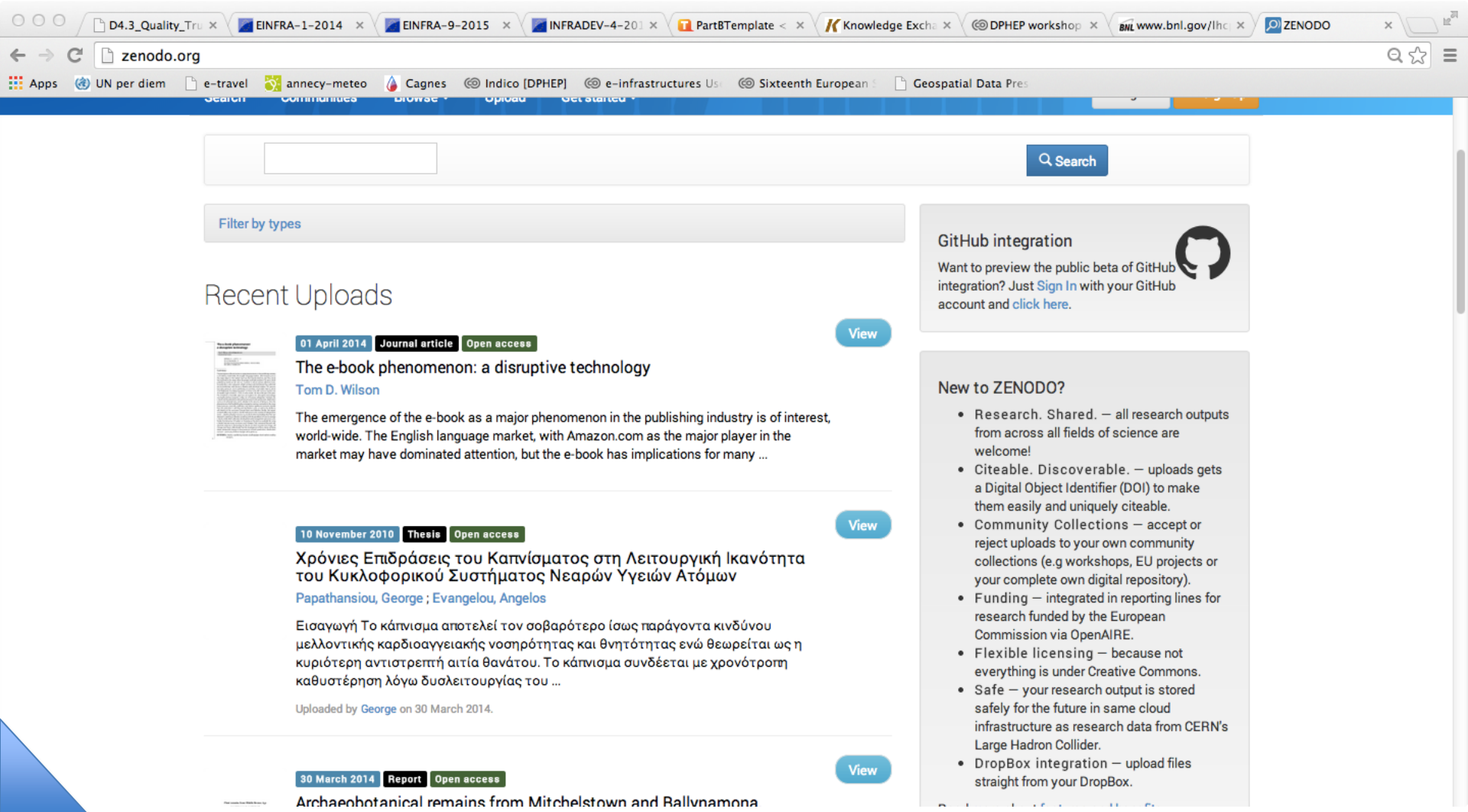
Guidelines Relating to Data Producers:

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

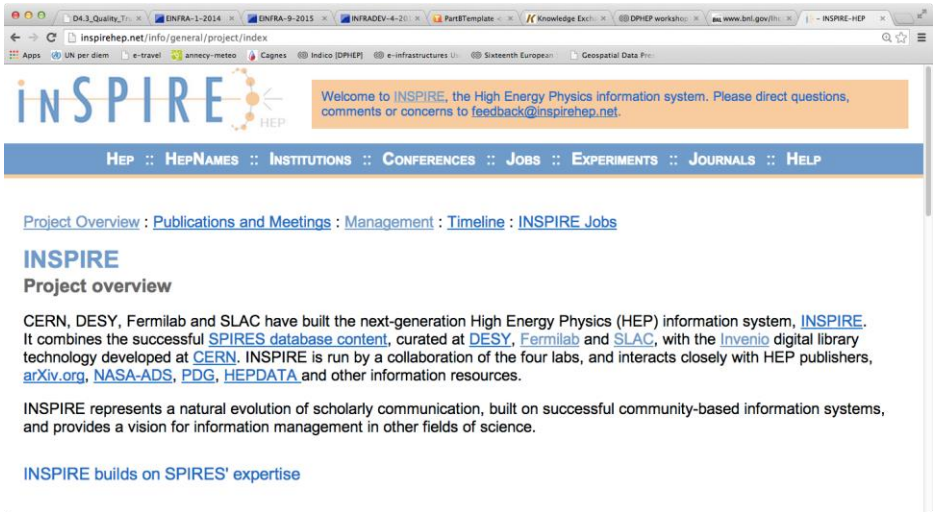
2. **Digital library** tools (**Invenio**) & services (**CDS, INSPIRE, ZENODO**) + domain tools (HepData, RIVET, RECAST…)
3. **Sustainable software**, coupled with advanced **virtualization** techniques, "snap-shotting" and **validation** frameworks
4. Proven bit preservation at the 100PB scale, together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

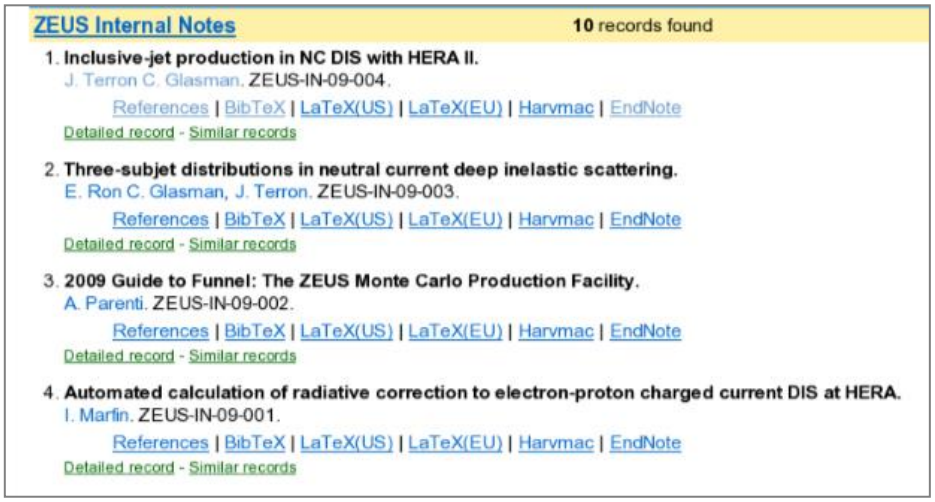
DPHEP Portal – Zenodo like?



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)





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”!