Long-Term D + K Preservation: Opportunities and Challenges

Jamie.Shiers@cern.ch

JRC-CERN Collaboration Workshop

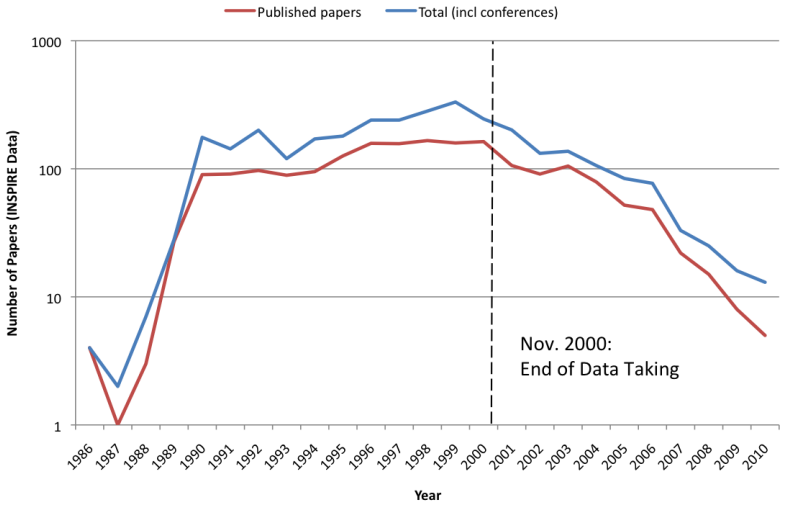
January 2014

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

Overview

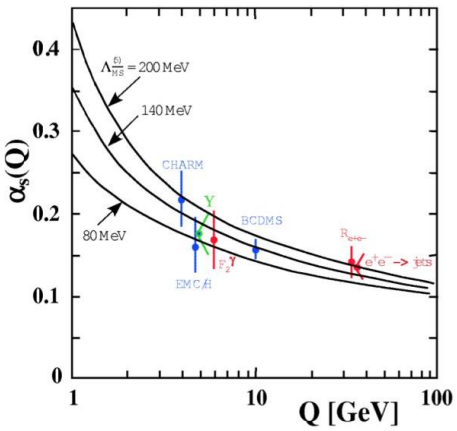
- Long-term, exa-scale data and knowledge preservation
- Involvement in the Research Data Alliance
- E-infrastructures User Forum and plans for INFRASUPP-7-2014

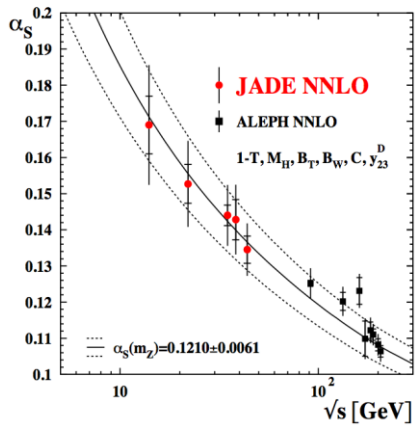
1 – Long Tail of Papers



3

2 – New Theoretical Insights



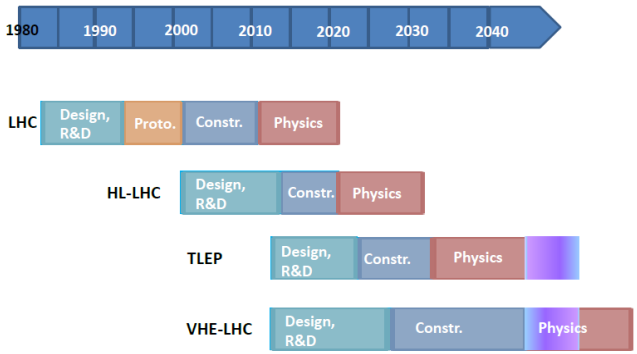


4

3 – "Discovery" to "Precision"

TLEP

possible long-term time line



Zimmermann

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)

1. DPHEP Portal

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

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

1. DPHEP Portal

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

1. DPHEP Portal

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

1. DPHEP Portal

Summary

1. **DPHEP portal: build in collaboration with other disciplines, including RDA IG**
2. **Digital libraries: continue existing collaborations**
3. **Sustainable "bit preservation" – certified repository as part of EINFRA-1-2014**
4. **"Knowledge capture & preservation": BIG CHALLENGE not addressed in multi-disciplinary way: next funding round?**
5. **Open "Big Data": a Big Opportunity (for RDA?)**

RDA Involvement

- CERN is involved in numerous Working and Interest Groups
- I am a member of the Technical Advisory Board and (Interim) Organisational Assembly
- See RDA as a valuable way of sharing knowledge and challenges with other disciplines and projects world-wide
- RDA outputs will help to achieve “Open Data” and other key goals

Postscript

- By far the greatest progress that we have made in “solving” long-term data preservation has been the result of **collaboration**

- We believe that we have a lot to **offer** other disciplines – as well as much to **benefit**