

Long-term preservation for HEP sessions summary

- Session summary :
 - Certification: Motivation, Benefits and Status
 - Collaboration Across HEP and with Other Disciplines: An EIROForum Sub-Group
 - Archiving and Preservation "in the Cloud"
 - CERN Analysis Preservation Portal
 - Risk management in LTDP
 - CWP Roadmap, Preparing Input to update of European Strategy for Particle Physics (and others elsewhere in the world)

Certification – Summary (HT Jamie Shears)

- ISO 16363 certification of CERN under way: covers end-to-end: from deposit to re-use.
- Depositors and (re-)users can ask any question, based on >100 agreed metrics
- Some sort of certification expected to become semi-mandatory on Run3 timescales
- Important that large facilities / projects on leading edge in this respect

Collaboration – Summary

- Space agency(-ies) defined OAIS
- Now coming to HEP to share (our) experiences
- Opportunities to expand this to other large European organisations, ESFRIs and possibly wider in the world
- *Didn't we do well?*

Cloud Archiving – Summary

- ESRIN outsources their “archive”: 20PB end 2017, 100PB within 10 years (*get/put i/f*)
- Base services not discipline specific – but some institutes may wish to maintain some control
- Combine e.g. Zenodo with a commercial archive (TDR)? “Solve” large storage requests to CERN
- Retain control of the “tricky bits” whilst out-sourcing the simpler ones?

Analysis Preservation – (HT Sunje Dallmeier-Tiessen, Mike Hildreth)

- An effort to provide a platform for the preservation of the entire workflow of an analysis.
- Re-using “maintstream” components (K8S, CWL, Invenio, ElasticStack, etc)
- Very encouraging collaboration across disciplines - preservation of the community as well as the data seems possible.

- REANA (Reusable ANALyses)
 - <http://reana.readthedocs.io> @reanahub
- Personal note – Exciting for publishing *all* the work of students/postdocs/etc – DOIs for everything, linked data
- Much of the infrastructure is already in place
- Many communities really need this

Risk Management

- Demonstrated by outage to CMS OpenData caused by incompatibility between SSL ciphers on certain clients and server
- Address with a consistent methodology - ISO 31000 – provides guidelines on principles and implementation of risk management
- Quantify Risk : the effect of uncertainty on objectives = Likelihood x Impact → prioritise.
- Not about *avoiding* risk, but mitigating it.
- Risk applies not only to data loss, but also to knowledge loss.

Data, software and analysis preservation

Challenges - (HT Mike Hildreth)

- **Longevity of executables stored in containers**
 - Directly related to rapid evolution of heterogeneous cpu architectures
 - Is there a set of standards that can guarantee future compatibility
- **Software sustainability**
 - Is it possible to make most algorithms architecture-agnostic?
 - Automatic optimization based on lower-level tools?
- **How to deal with “big-data” type analyses in terms of preservation?**
 - Data is not “reduced”: entire dataset is queried, could vary in time
 - Algorithms are site installation specific: how to preserve without whole site?
- **Lack of preservation tools for ordinary physicists**
 - Workflows not automatically captured, annotated

Roadmap

- Near term: (1 year)
 - Demonstrate a fully functional version of the current container solution (CERN Analysis Portal) and orchestration of multiple containers for HEP Production executables [+Workflow & Resource Management, Distributed Computing WGs]
 - Investigate the limitations of the container solution
 - Study use cases for analysis workflows
 - Investigate if this approach will enable types of analyses currently not possible. And types that we can't. Interplay with BigData-analytic-style solutions

Roadmap

- Medium term: (3 years)
 - Development of prototype analysis system(s), tools, embedding preservation elements [+Data Analysis and Interpretation WG][engage DevOps community?]
 - Understand the full range of things that can and can't be captured
 - Research on things that can't be captured
 - Investigate the limitations of the container solution (evolving)