Data, Software, & Analysis Preservation: Overview of CWP Conclusions

Mike Hildreth, with much help from the DHEP Community

Overview

CWP Content developed over 2017 HSF workshops:

- Meetings at San Diego, Princeton, Annecy, Seattle
- CWP Chapter outline and text shared via Google docs until final version was prepared
- ~15 pages, attached to Indico agenda page
- Includes R&D roadmap, emphasis on links to other HSF activities
- Emphasis on re-use (mostly internal) of data, software, and analyses
  - FAIR principles
- Highlights…

(Some) Challenges

- 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

Some of these obviously overlap with other WGs

R&D 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.
- 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)

Other obvious overlaps with Data Access, Organization, and Management WG, Software Development WG, Event/Data Processing Frameworks WG

R&D Roadmap, continued

- Longer term: (5 years)
  - Tools developed for anyone to preserve an analysis that they are starting
  - Tools developed for “large” executable preservation

Comments:

- Preservation constraints will likely not drive future developments,
- BUT, preservation can be integrated into new developments with adequate coordination and planning, especially if preservation techniques can
  - Provide added capability for analysts
  - Be a natural extension of distributed computing models
  - Be incorporated into software and workflow development tools

Collaboration/Coordination with other groups will be very important