DASPOS: Common Formats?

Mike Hildreth

Université de Notre Dame du Lac & Fermilab

Representing the DASPOS Project

Quick DASPOS Overview

- Data And Software Preservation for Open Science
- multi-disciplinary effort recently funded by NSF
  - Participants: Notre Dame, University of Chicago, University of Illinois Urbana-Champaign, University of Nebraska Lincoln, New York University, University of Washington, (BNL, FNAL)
  - Open communication and advice from CERN, OSG, DPHEP, DataNet, etc.
- Links HEP effort (DPHEP+experiments) to Biology, Astrophysics, Digital Curation
  - Diverse set of participants
- aim to achieve some commonality across disciplines in
  - meta-data descriptions of archived data
    - What's in the data, how can it be used?
  - computational description
    - how was the data processed?
      - i.e.: follow Tier 3 reconstructed data to final physics result
  - impact of access policies on preservation infrastructure

DASPOS Overview II

- In parallel, will build test technical infrastructure to implement a data preservation system
  - "scouting party" to figure out where the most pressing problems lie, and some solutions
    - incorporate input from multi-disciplinary dialogue, use-case definitions
  - Will translate needs of analysts into a technical implementation of meta-data specification
  - Will create means of instantiating computation from metadata description
  - Will implement "physics query" infrastructure across small-scale distributed network
  - end result: "template architecture" for data preservation systems
    - Translation: we will attempt to build something that works for "Level 3" data preservation and see how far we get

Year 1: Workshops

Plan two Workshops on Data/Software/Analysis Preservation in 2013:

1. HEP-Focused (Spring @ CERN)
   - Address issues of commonality (or lack thereof) in HEP D/S/A preservation across the HEP community
   - Focus on "Level 3" data tier
   - Can we agree on a preliminary set of use cases for re-analysis?
   - Can we agree on a preliminary set of descriptors/metadata that can be used to characterize:
     - The analysis performed?
     - The computation steps used to produce the final result?
     - The software required to produce the final result?
   - One "Level 2" data question: can we agree on a common 4-vector format and descriptors?
     - May be an easier place to start…

Year 1: Workshops

2. Multi-Disciplinary D/A/S Preservation "Survey" Workshop
   (~ Summer, Satellite of major Data Preservation conference)
   - Obtain overview of D/A/S workflows in other scientific disciplines using large datasets
     - Already have contacts with Astro, Bio-Informatics, etc.
   - Intent: attempt to define a level of commonality for
     - Metadata descriptions of data, processing, software
       - Can we re-use ones sufficient for HEP with some adaptation?
     - At a base level, can we create a common framework?
   - Explore: impact of access policy decisions on
     - Storage architectures/networking
     - Content of metadata
     - Implications for HEP?
       - Interfacing with OAIS?

Brief Conclusions:

- First, please help in this effort!
  - We must coordinate with other D/A/S preservation efforts
  - Interface with DPHEP is critical
  - Reminder: not trying to solve the world's problems
    - A "scouting effort"
- Second, take away the message that we are working towards common descriptions a higher level of abstraction than some current efforts
  - Start to think about how HEP data descriptions and structure can fit within common formats