DASPOS Data and Software Preservation for Open Science

National Science Foundation
WHERE DISCOVERIES BEGIN

Introduction to Data/Analysis Preservation Discussion

K. Bloom, R. Gardner, M. Hildreth, E. Long, R. Johnson, M. Neubauer

For the DASPOS Team

Introduction to DASPOS

- Data And Software Preservation for Open Science
  - multi-disciplinary effort recently funded by NSF
  - Notre Dame, Chicago, UIUC, Washington, Nebraska, NYU, (Fermilab, BNL)
- Links HEP effort (DPHEP+experiments) to Biology, Astrophysics, Digital Curation
  - includes physicists, digital librarians, computer scientists
  - 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 (ontology development)
      - how was the data processed?
      - can computation replication be automated?
  - impact of access policies on preservation infrastructure

DASPOS Structure

- In parallel, will build test technical infrastructure to implement a knowledge 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, policy discussions
  - Will translate needs of analysts into a technical implementation of meta-data specification
  - Will develop means of specifying processing steps and the requirements of external infrastructure (databases, etc.)
  - Will implement "physics query" infrastructure across small-scale distributed network
- end result: "template architecture" for data/software/knowledge preservation systems

DASPOS Plans

- Three "Fact-Gathering" Workshops in 2013:
  - HEP-centric (This Workshop):
    - Can experiments agree on the types of data they would like to preserve?
      - software and analysis preservation, too
    - Can we begin to define some global metadata?
  - Multi-Disciplinary: (Fall, maybe at "Big Data" in October)
    - What are problems, use cases in other fields? (Astro, Bio, etc.)
    - What is the commonality between these and HEP?
      - can we think about common infrastructure?
  - Technical: (TBD)
    - Survey of archival architectures
    - learn from infrastructure work developed for other problems
    - try not to re-invent multiple wheels…

This Session

- Our aim was to get an overview of how different experiments do analysis:
  - What data is required?
  - How many steps of processing are there?
  - How is all of this documented?
  - Is the analysis software archived?
  - …
- What we are trying to understand is if we can derive a common "data model" for data use and analysis description that might allow people to describe and archive what they did, and do so in a reproducible way
  - purely a fact-gathering exercise
- Another issue: Use cases for data & analysis preservation:
  - What is important for you and your experiment?
  - How might data be used in the future?

Additional Information

List of Questions:

- Disaster Recovery (if so, how fast backup and what is acceptable loss, 1 hour, 1 day, 1 week, etc)
- Reuse
- Reproduce or Recreate (i.e. reimplement algorithms)
- Education
- Short term access, long term access
- Emulation vs. Migration
- Future Research
- Activitives for raw data, analytical data, algorithms, software, environment (i.e. OS)
- Which activities and what is important to preserve and share in
  - < 2 years
  - < 5 years
  - < 10 years
  - < 20 years