DASPOS Overview

Mike Hildreth

Université de Notre Dame du Lac & Fermilab

Representing the DASPOS Project

Introduction

- Data And Software Preservation for Open Science
- Multi-disciplinary effort recently funded by NSF
  - 3 Years of funding
  - 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 other disciplines
  - Close collaboration with national labs, experimental efforts
- Diverse set of participants:
  - HEP, including computation experts
  - Computer Scientists, including HPC experts
  - Digital Librarians, including SDSS curators, Bioinformatics

Overview

- Goal: explore common solutions to Data and Software Preservation
- 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
- Two aspects of the project: "T" shape
  - Discovery & Coordination Activity
    - workshops to address relevant issues, communication
    - broad interest and exploration
  - Prototyping & Experimentation Activity
    - building test infrastructures, computation/curation challenge
    - focus on building a vertical slice of architecture

Discovery & Coordination: Year 1 Workshops

Plan 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
   - Familiarize outside experts with HEP problems
   - One "Level 2" data question: can we agree on a common 4-vector format and descriptors?
     - May be an easier place to start…
   - 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?

Discovery & Coordination: 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
     - Ontology development: Metadata descriptions of data, processing, software
       - Can we re-use ones sufficient for HEP with some adaptation?
       - rely on expertise from Digital Librarians
     - 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?

Discovery & Coordination: Year 1 Workshops

3. Data Model and Query Symantics

   (~ late 2013, FNAL or CERN?)

   – CS and infrastructure specialists meet HEP head-on

   – Once initial ideas on metadata have had a chance to mature and some agreement has been reached on what should be stored,

     - can we achieve a common logical model for the organization of the stored data?
     - can we arrive at a set of ways in which the metadata will be queried?
       – how are future analysts going to look at this data, and what organizational structure does that imply?

   – Input from this meeting, plus technical results from prototyping, will help drive the architecture

Discovery & Coordination: Year 2 Workshops

Three more workshops in Year 2 to round out the topics:

1. Software Sustainability
   - establishment of software life-cycle and needs
   - long-term models and mechanisms for software preservation
2. Preservation Policies
   - explore access and preservation policies and the constraints these impose on data preservation
3. Technical Developments in Storage Architectures
   - collect best practices in large-scale and long-term storage architectures

Prototyping & Experimental Task

- In parallel, will build test technical infrastructure to implement a data preservation system
  - 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
  - product: "template architecture" for data preservation systems

Definition of "success":

We will consider this project a success if we have created a clear intellectual structure and useful prototypes that enable others to carry the effort forward with greater resources

Prototyping & Experimental Task

- Technical infrastructure portion of the project is based on building prototype systems with the desired functionality
  - "scouting party" to figure out where the most pressing problems lie, and some solutions
    - incorporate input from multi-disciplinary dialogue, use-case definitions
- Primary task #1: Create Data Model & Query Semantics
  - What data is to be stored, how is it queried?
    - information from Workshops
  - Test these models by building small-scale systems
  - pay particular attention to scalability issues
  - Test effects of distributed data on query specification

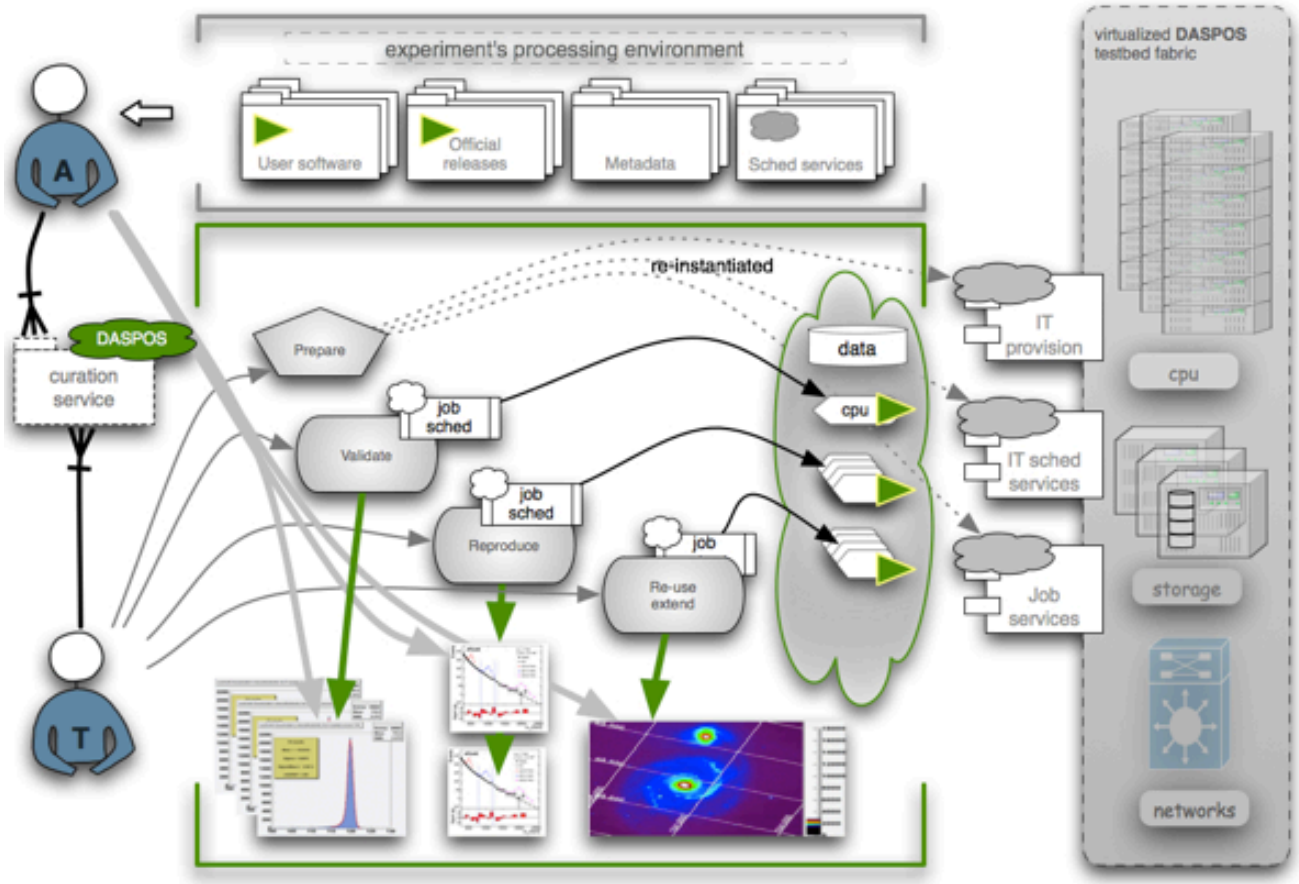
Prototyping & Experimental Task

- Primary Task #2: Define Elements of Software Reproducibility
  - How do you re-create an analysis result?
  - Several elements:
    - define naming scheme for user specification of configuration
    - dependency analysis to ascertain which software is needed
    - create means of "task insertion" = execution based on the provided information
    - provide reproducibility checking & validation for results

- Small-scale "Reproducibility Challenges" will be conducted to track progress, insure that necessary elements and correct descriptions are present

Curation Challenge

- Test analyst “T” will reproduce an analysis at a level specified by and Audit Team “A”
  - Final Demonstrator of DASPOS prototype architecture



M. Neubauer, UIUC

UNIVERSITY OF NOTRE DAME Fermilab

Outlook:

- Broad, Ambitious Effort
  - relatively small team, but diverse set of expertise
  - Reminder: not trying to solve the world's problems
    - A "scouting party": push forward in what looks like a good direction without worrying about full world-wide consensus
- Coordination is Key
  - We must coordinate with other D/A/S preservation efforts
  - Interface with DPHEP is critical
  - knowledge from previous work, other fields essential

Please help if you can!