Data and Software Preservation for Open Science (DASPOS)

- ***Preservation in this context:*** "***Ensuring the continued usability of the data and software necessary to conduct science.***"
- Project to
  - provide appropriate **data, software, and algorithmic preservation for HEP**
  - Address importance of preserving the contexts necessary to understand, trust, and re-use data
- Proposal, written under leadership of M. Hildreth (ND), was submitted to NSF/MPS as a PIF
  - Participants from HEP (LHC and Tevatron), Digital Libraries and Computer Science
  - Endorsed by OSG (initial white paper by Ruth)
  - Funded for 3 years at requested level (details to follow)

Broader Impact

- Archiving of HEP data may require some HEP-specific technical solutions
- The broader activity of curating data includes
  - data management, appraisal processes, policy development, and discovery and access services
  - will have commonalities with other disciplines

➢ DASPOS will create a template for preservation

  - useful across many different disciplines, leading to a broad, coordinated effort

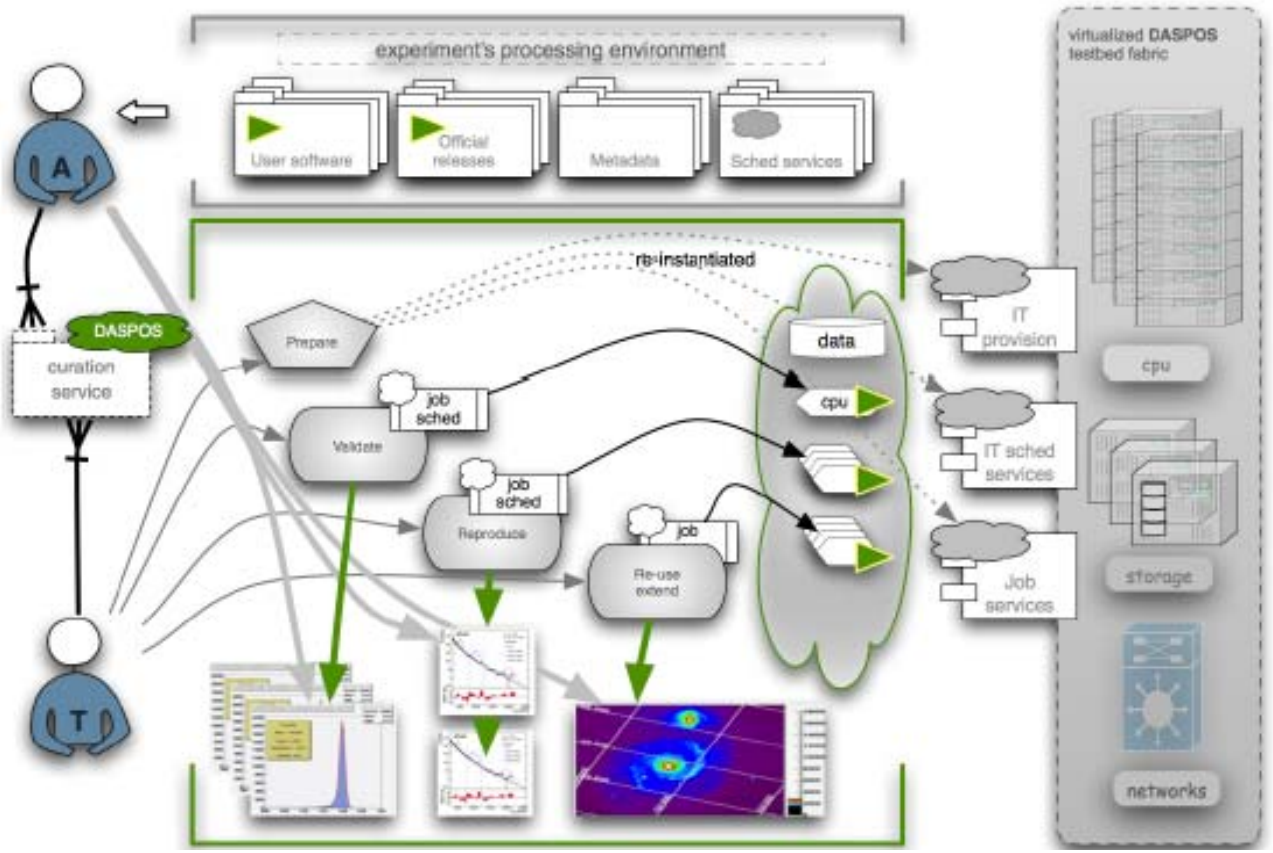
The Project …

- … is divided into two distinct but mutually-informing activities
  - The **discovery and coordination activity** will bring together a broad community of experts and stakeholders to define, discuss, and document the details of data and software preservation for high energy physics and other fields. This activity will be organized as a series of highly structured public workshops.
  - The **prototyping and experimenting activity** will address two essential problems in preservation by constructing prototype software and gaining experience through experimentation. The two key areas of research are the data and query models and the software sustainability models.

Goal of DASPOS

- “scout out” solutions to the most pressing technical problems, and make them available to those constructing preservation systems
  - Establish a dialogue with other fields facing preservation and re-use issues with Big Data.
  - Identify areas of commonality and outline where solutions diverge due to specific needs.
  - Develop metadata to support the preservation and re-use of HEP data, and its related software and computational algorithms. Design the metadata so as to meet the needs of as many other fields as possible for wide re-use.
  - Define a reference architecture for a data preservation system targeted for HEP but coordinated with other fields. Include decision points where policy choices impact the architectural structure.
  - Develop a preservation validation test-bed on which a technical implementation of the reference architecture can be developed and constructed.
  - Perform a *Curation Challenge*, where a physics data analysis is conducted based solely on curated and archived data.

Curation Challenge



- The auditing team (A) prepares the prototype DASPOS curation service with necessary semantic description of the tasks, software and processing environment; a technical execution team (T) uses these data to marshal the processing environment from the DASPOS testbed fabric, and performs successively a validation, reproduction, and an extended analysis which is later audited for accuracy by the A-team.

Management Plan

- Project Coordination
  - PI Mike Hildreth (Notre Dame)
  - Co-PI Mark Neubauer (UIUC)
  - Kyle Cranmer (NYU)
- Data and Query model
  - Co-PI Doug Thain and Co-PI Jarek Nabrzyski (ND)
- Software and processing sustainability
  - Co-PI Rob Gardner and Bob Grossman (UC)
- Coordination of Curation Challenge
  - Co-PI Mark Neubauer and Rob Gardner

Prototyping and Experimentation Activity

Year 1	**D10**	Data Model and Query Semantics Document	Month 12
	D11	Prepare a testbed in support of Reproducibility Challenges	Month 6
	M1	Local Implementation of Query Prototype	Month 12
Year 2	**D12**	Virtualized DASPOS testbed fabric and associated services necessary to support the above Curation Challenge milestones	Month 16
	M2	Demonstration of Reproducibility of analysis result use-case using maximally virtualized environment construction	Month 24
	M3	Distributed Implementation of Query Prototype	Month 24
	M4	Demonstration of Reproducibility of analysis result use-case using source-based environment construction	Month 24
Year 3	**D13**	Open source software that provides a reference implementation of the abstract data model and query language	Month 36
	D14	Prepare a prototype method for retrieving data and processing semantics for identified use-cases	Month 36
	M5	Use of DASPOS semantics and reference architecture to capture an end-to-end data, software, processing environment and challenge task including validation and reproducibility criteria	Month 27
	M6	Use of DASPOS semantics and reference architecture to validate and reproduce the curation challenge task	Month 30
	M7	Use of DASPOS semantics and reference architecture to re-used and extended preserved data in the context of the Curation Challenge	Month 36

Participation and Funding

- Institutions getting Funding
  - Notre Dame, U of C, UIUC, UNL, UW
    - ND: part of Computing Director, 0.5 FTE Comp Scientist, 0.5 FTE programmer, 2.5 yr x 2 grad stud.
    - UofC: part of Rob, 2 yr of Comp Scientist, 2.5 yr grad student
    - UIUC: 2.5 yr grad student
    - UNL: 2.5 yr grad student
    - UW: 3 yr grad student
  - Sustained total of 2 FTE Comp Professionals and 5 FTE students
  - Funding starts in FY 2012 (~now)