Data and Knowledge Preservation (in High Energy Physics)

Mike Hildreth
University of Notre Dame
representing the DASPOS Project

DASPOS Data and Software Preservation for Open Science

Data preservation in HEP

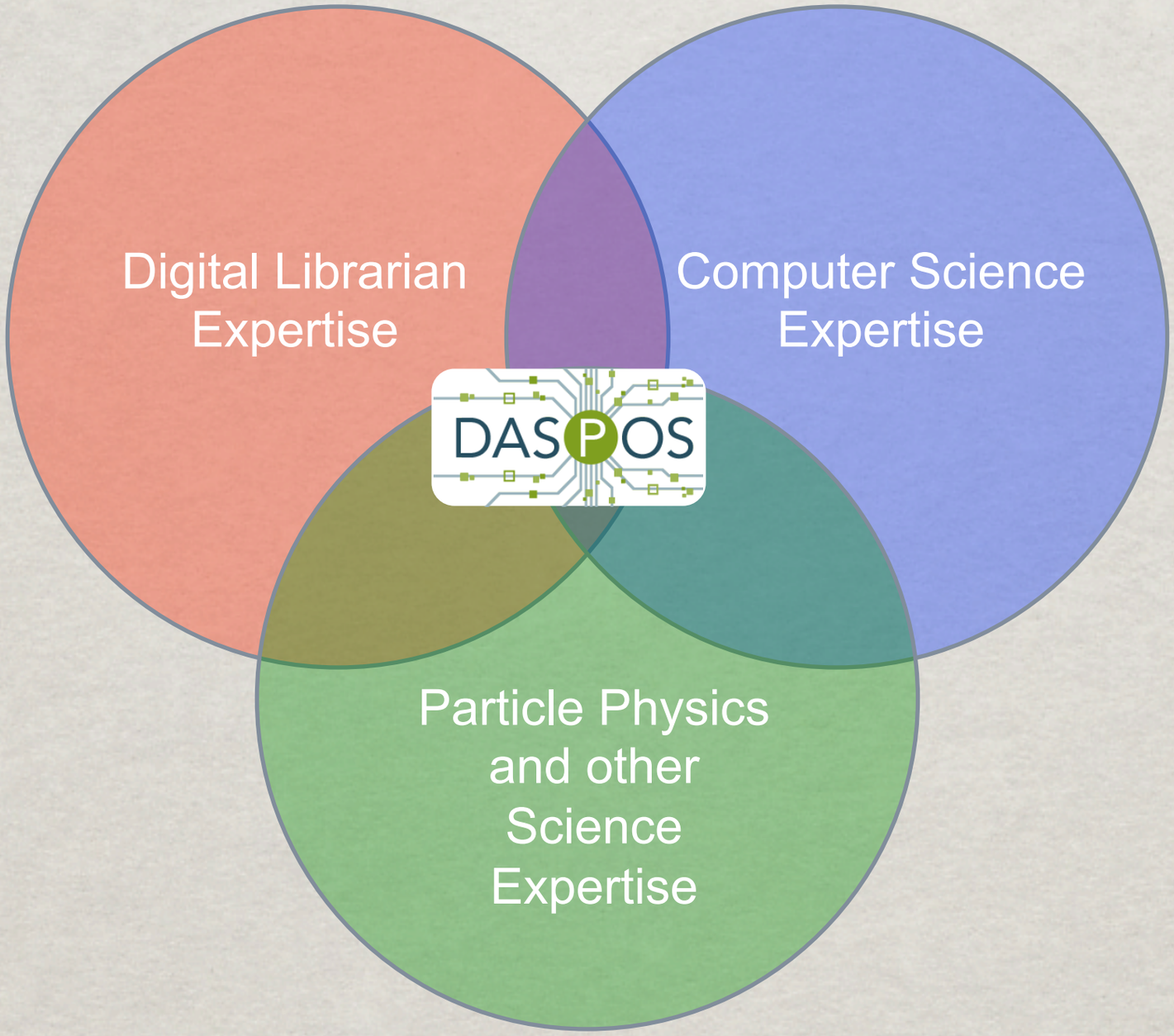
- What to do with 100s of PB of data?
  - Irreplaceable resource
  - should be preserved, some how, for the future
- DPHEP Working Group
  - Convened by International Committee on Future Accelerators (ICFA)
  - ~ 100 members from different HEP experiments, Labs
  - Two Reports:
    - DPHEP-2009-00, http://arxiv.org/pdf/0912.0255
    - DPHEP-2012-01, May 2012, arXiv:1205.4667v1
  - Conclusions:
    - "an urgent and vigorous action is needed to ensure data preservation in HEP"
    - "A clear and internationally coherent policy should be defined and implemented"

DASPOS

- Data And Software Preservation for Open Science
  - multi-disciplinary effort, funding started 9/12
  - Notre Dame, Chicago, UIUC, Washington, Nebraska, NYU, (Fermilab, BNL)
- Links HEP effort (DPHEP+experiments) to Biology, Astrophysics, Digital Curation, and other disciplines
  - 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/metadata development)
      - how was the data processed?
      - can computation replication be automated?
    - impact of access policies on preservation infrastructure

DASPOS Overview

- How to catalogue and share data
- How to curate and archive large digital collections
- Ontology/ Metadata expertise



- How to build databases and query infrastructure
- How to preserve software and functionality
- How to develop distributed storage networks

- What does the data mean?
- How was it processed?
- How will it be re-used

DASPOS Activities

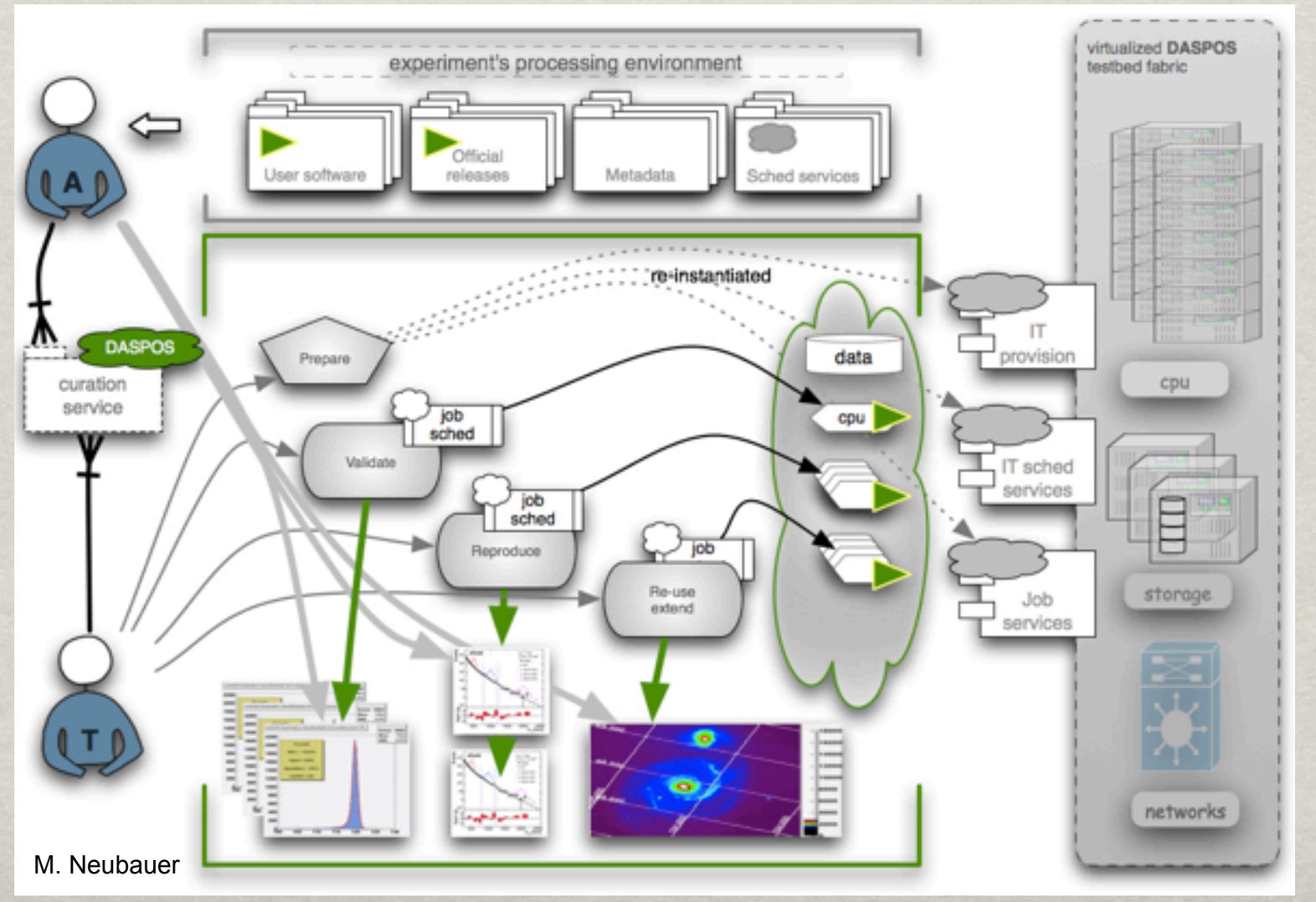
- Three "Fact-Gathering" Workshops in 2013:
  - HEP-centric (March, at CERN, joint with DPHEP):
    - Can experiments agree on the types of data they would like to preserve?
      - software and analysis preservation, in addition
    - Can we begin to define some global metadata?
  - Multi-Disciplinary (July, Indianapolis, at JCDL):
    - 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 Infrastructure (TBD):
    - Survey of archival architectures
    - learn from infrastructure work developed for other problems
    - try not to re-invent multiple wheels…

DASPOS Activities

- 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.)
    - automatic instantiation of workflows?
  - Will implement "physics query" infrastructure across small-scale distributed network
  - End result: "template architecture" for data/software/knowledge preservation systems

DASPOS

- Final Milestone: “Curation Challenge”
  - an analyst will reproduce some physics result using only curated information
  - success defined by external auditing team



M. Neubauer

Workshop I

Establishment of Use Cases for Archived Data and Software in HEP

- **Attendees:** Participants from all of the HEP experiments considering long-term data preservation and access issues (4 LHC experiments, BaBar, D0/CDF); joint with DPHEP
- **Organizers:** A team consisting of the digital librarians from University of Chicago and Notre Dame and HEP physicists from Notre Dame, University of Chicago and University of Illinois at Urbana-Champaign
- **Location:** CERN
- **Purpose:** (i) Establish use cases for data access and re-use, especially for the larger DPHEP data tiers, since this will be a primary driver of the preservation architecture, (ii) define what data and associated information supports the use cases, and (iii) identify a preliminary set of metadata that would serve the needs of the HEP community in accessing the various forms of archived data/algorithms.
- **Inputs**: Questionnaire based on Data Curation Toolkit addressing: use-case scenarios for data re-use and archiving including intended audiences; current practices and policies for data use; data types; and high-level description and rights metadata necessary for discovery and access.

Workshop I

Other Agenda Topics

- prospects for commonality in outreach formats
  - issues being driven by CMS outreach program in Finland that will release "real" data
  - open question, limited by available people
- high-level analysis preservation
  - HepData ⇔ Rivet ⇔ Theorists' Analysis Archive
    - Can an extension of the HepData/Rivet infrastructure serve as a common platform for high-level analysis preservation?
    - Now discussing Rivet back-end interface for RECAST (NYU/UNL)

Outcomes

- Analysis of commonality in data processing, analysis chains
  - first look at abstraction of workflow steps
    - still looking at how best to represent this so that it will be useful for other disciplines

Workshop II

Survey of Commonality with other Disciplines

- **Attendees**: Broad participation from many NSF supported science efforts.
- **Location:** Satellite workshop at IEEE/JCDL (Joint Conference on Digital Libraries) in Indianapolis, IN, July 2013.
- **Organizers:** Digital librarians, HEP, and Computer Scientists
- **Purpose:** (i) Explore areas of commonality and difference, (ii) identify common metadata standards that could be designed to allow generic access and indexing of cross-disciplinary research data, and (iii) identify cross-disciplinary services that would support data preservation (e.g. software repositories).
- **Inputs:** A panel discussion with many cross-disciplinary participants; break out sessions targeting sub-topics of interest within data preservation.

Workshop II

DASPOS

DASPOS Data and Software Preservation for Open Science

www.daspos.org

ABOUT | PEOPLE | WORKSHOPS | RESEARCH | REPORTS

DASPOS Workshops

- Use Cases for Archived Data and Software in HEP
- **Commonality with other Disciplines**
- Data Model and Query Semantics
- Software Sustainability
- Preservation Policy
- Technical Storage Architectures

WORKSHOP 2

Survey of Commonality with other Disciplines

Date: Thursday, July 25, 2013

Agenda: Click here for schedule of events

Purpose: (i) Explore areas of commonality and difference, (ii) identify common metadata standards that could be designed to allow generic access and indexing of cross-disciplinary research data, and (iii) identify cross-disciplinary services that would support data preservation (e.g. software repositories).

Inputs: A discussion framework similar to that of Workshop 1 will be developed and will also be used to conduct individual or small group discussions with targeted colleagues not available for the workshop (e.g. the research staff involved in archiving the Sloan Digital Sky Survey-II).

Outcomes: Provide extensive information about preservation efforts in other disciplines.

Panel Discussion Video: Click here for video recording of panel discussions

Round Table Discussion Video: Click here for video recording of the round table discussions

Discussion: Round Table Discussion Notes

Workshop II

Panel Participants

Dr. George O. Strawn
Director, National Coordination Office (NCO)
National Science Foundation

Dr. Reagan W. Moore
Director, Data Intensive Cyber-Environments
Chief Scientist, RENCI
Professor, School of Information and Library Science
University of North Carolina at Chapel Hill

Dr. Chris Mattmann
Senior Computer Scientist
NASA Jet Propulsion Laboratory

Dr. Don Petravick
Principal Investigator
Dark Energy Survey Data Management System
National Center for Super Computing Applications
University of Illinois at Urbana-Champaign

Dr. Matthew Mayernik
Research Data Services Specialist
NCAR Library / UCAR Integrated Information Services
National Center for Atmospheric Research (NCAR)
University Corporation for Atmospheric Research (UCAR)

Prof. Michael Witt
Associate Professor of Library Science
Interdisciplinary Research Librarian
Purdue University

Dr. Micah Altman
Director of Research
Head/Scientist, Program on Information Science
MIT Libraries

Dr. Clifford Lynch
Executive Director
Coalition for Networked Information

Dr. Line Pouchard
Information Scientist
Scientific Data Group
Oak Ridge National Laboratory
US Department of Energy

Workshop II

Round-Table Topics:

- Policy based Data Management
- Reuse of Big Data, complex digital objects & Scientific Workflows
- Software & Algorithmic Preservation for Open Science

Outcomes:

- Common themes:
  - Provenance of data, Workflows, definition of workflow, reproducibility
  - Software preservation
  - Policy based data management
  - Metrics, citations
  - Economics
- Other tools/concepts to explore (understand uses in other disciplines):
  - Taverna, MyExperiment, iRODS, etc.

Technical Advances

Software/Workflow/Environment Preservation (Chicago)

- Prototype Workflow capture/validation on one step in the analysis chain
  - “Slim/Skim” step in ATLAS environment
  - Start with actual ATLAS analysis workflow and data stored at a Tier 2
  - capture user workflow process in service database, validate against existing results
  - Next:
    - Execute using existing virtualized execution infrastructure at the ATLAS Midwest Tier 2 Center
    - Deploy captured code, data and environment on OpenStack managed preservation platform
    - Develop an "Is the data still alive” monitoring infrastructure

DØ Effort (Washington)

- Establishing analysis platform on virtual machines outside of FNAL infrastructure
- have one VM running SL6, second as a cvmfs server to provide DØ software
- experience will translate to LHC efforts

Technical Advances

DASPOS

Data Model & Query Semantics (Notre Dame)

- Two new graduate students added to project this fall
  - pilot projects aimed at developing an understanding of the HEP data model, including processing steps, environment
  - weekly “DASPOS workshops” with HEP/Computing/CS experts
    - cross-fertilization of ideas, vocabulary

- Early project goal:
  - demonstrate a working prototype of the data model and query language on a fixed data set stored on a local filesystem independent of the LHC infrastructure

- REU Projects:
  - “Data git”: ontology/metatdata designed to track changes in the processing state for a given file/repository
  - CMS computing environment catalogue: offshoot of cyber-infrastructure project to provide computing interoperability for generic Condor nodes

Conclusions

Internal Challenges:

- Many, many things to learn
  - merely just figuring out a common vocabulary took a couple of months
  - integration of different points of view and expertise is critical, but it doesn't happen quickly

External Challenges:

- Coordination
  - tens of data preservation efforts scattered across the globe in different disciplines; growing realization that "knowledge preservation" is the real issue
    - at the highest levels, all have different requirements
    - at the most basic levels, we are all doing the same thing
      - arguably, most other disciplines are "less complicated" than HEP
      - a framework with sufficiently generic pieces should be able to be adopted in a simplified manner by other disciplines
  - Would like to see joint workshops (DoE/NSF/NASA -sponsored) on these issues
    - International partners? Role of RDA?