Data and Software Preservation for Open Science (DASPOS)

Summary of Workshop 2: Interdisciplinary Commonalities

July 25, 2013

Held jointly with Digital Preservation of Research Methods and Artefacts workshop at ACM/IEEE JCDL, Indianapolis, Indiana

R. Johnson, E. Long, N. Meyers, DASPOS co-organizers, with
R. Gardner, M. Neubauer, J. Nabrzyski, D. Thain, C. Vardeman and …
DASPOS/DPRMA attendees

Overview

This workshop held jointly with the 1st International Workshop on the Digital Preservation of Research Methods and Artefacts (DPRMA 2013) and continued the DASPOS exploration of the key technical problems that must be solved to provide data and software preservation for open science. During this workshop, under the aegis of the DASPOS project, we conducted a discovery and coordination activity to bring together a broad community of experts and stakeholders to begin to define, discuss, and document the details of data and software preservation across different disciplines. The workshop aimed through panel presentations, discussion with panelists, and round table discussions, to explore appropriate data, software and algorithmic preservation approaches, including the contexts necessary to understand, trust and reuse data.

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).

Summary of Workshop Activities

The DASPOS Multidisciplinary Commonalities workshop was co-located with DPRMA 2013 at the ACM/IEEE joint conference on digital libraries (JCDL) .

DPRMA workshop paper presentations were organized by Kevin Page, David DeRoure, Andreas Rauber, and Jun Zhao. The papers re available in the proceedings available from the ACM DL at http://dl.acm.org/citation.cfm?id=2499583 and brief citations for the papers appears below:

This work was supported by grant NSF-PHY-12473

Data and Software Preservation for Open Science (DASPOS)

1. Christophe Guéret : *Digital archives as versatile platforms for sharing and interlinking research artefacts* DOI: 10.1145/2499583.2499588
2. Raúl Palma, Oscar Corcho, Piotr Hotubowicz, Sara Pérez, Kevin Page, Cezary Mazurek : *Digital libraries for the preservation of research methods and associated artifacts* DOI: 10.1145/2499583.2499589
3. David De Roure : *Towards computational research objects* DOI: 10.1145/2499583.2499590
4. Tomasz Miksa, Andreas Rauber : *Increasing preservability of research by process management plans* DOI: 10.1145/2499583.2499591
5. Robert Sanderson, Herbert Van de Sompel, Peter Burnhill, Claire Grover : *Hiberlink: towards time travel for the scholarly web* DOI: 10.1145/2499583.2500370

DPRMA, like DASPOS is engaged by the process of research in both the sciences and humanities which has, and continues, to undergo significant change in addressing the needs of our ever more digital world. Researchers are adapting to the opportunities presented by working at scale with increasingly large datasets, creating methodologies and tooling for assistance and automation, and undertaking multi-disciplinary collaboration with colleagues and specialisations distributed around the globe.

This brings with it challenges for the capture, publication, and preservation of research output. In this world a single document or journal paper -- perhaps by a single author with a narrow subject focussed bibliography -- is no longer sufficient for useful encapsulation of the complete research output. This is particularly the case when considering the need to disseminate, reproduce and reuse methods and findings as the foundation of ongoing scholarly research and academic discourse.

DPRMA's papers presented at the workshop considered how Digital Libraries can adapt to meet these needs. Starting with the complex digital objects needed to store the multi-format artefacts such as datasets, workflows, results and publications, the workshop will discuss how they they be captured, stored, associated, retrieved, and visualised. Can, or should, Digital Libraries address the needs of scale presented by big data directly and wholly, or play a well-defined role within an ecosystem of interoperable services? What are the challenges for curation of dynamic resources often more akin to software than documents, where iterative experiments comprise of changing datasets, codes, and authors? What additional research context should be preserved in addition to traditional dissemination mechanisms? What models and semantics can capture this context, and what role can provenance, versioning, and dependency analysis play in their preservation? How will researchers access and reuse these preserved artefacts?

DASPOS Workshop Panels & Panel Presenters

DASPOS in its effort to explore areas of commonality and difference convened two panels at the workshop. The first panel was moderated by DASPOS PI Gordon Watts and featured:

- Chris Mattmann, Senior Computer Scientist NASA Jet Propulsion Laboratory

Data and Software Preservation for Open Science (DASPOS)

- Reagan Moore, Director of the Data Intensive Cyber Environments Center at UNC Chapel Hill
- Don Petravick, Principal Investigator Dark Energy Survey Data Management System at NCSA –U of I at Urbana Champaign
- George Strawn, Director of NITRD's National Coordination Office (NCO).

The second panel was moderated by DASPOS PI Doug Thain and featured:

- Micah Altman, Director of Research and Head/Scientist, Program on Information Science for the MIT Libraries
- Matthew Mayernik, Research Data Services Specialist , NCAR
- Michael Witt, Project director for the Purdue University Research Repository (PURR)

The above DASPOS Panelists' Presentations are available linked on the workshop webpage at: https://daspos.crc.nd.edu/index.php/workshops/workshop-2.

The DASPOS Panels were recorded and raw video of the panels is available at : https://www.youtube.com/v/_vK57JooR9c .

Round Table Topics

DASPOS in its effort to identify common standards and tools to facilitate cross-disciplinary research data, and identify cross-disciplinary services that would support data preservation convened three round tables at the workshop.

- Policy based Data Management lead off by Reagan Moore w/DASPOS PI Sr.Personnel Elisabeth Long, Matt Mayernik, Michael Witt, Benoit Raybaud, Charles Vardeman
- Reuse of Big Data, complex digital objects& Scientific Workflows lead off by Don Petravick and Micah Altman w/DASPOS PI Sr. Personnel Rob Gardner and Jaroslaw Nabrzyski and Clifford Lynch from CNI.
- Software & Algorithmic Preservation for Open Science lead off by Chris Mattman and George Strawn w/DASPOS PI Doug Thain

Round tables convened under their respective leaders and then met to provide a summary of their conclusions/highlights to the larger group. The DASPOS Round Table summary session was recorded and a raw video of the summary session is available at: http://youtu.be/xlyUClZ-Inc and the Round Table Discussion Notes are posted on the DASPOS workshop 2 webpage at: https://daspos.crc.nd.edu/images/daspos/workshops/workshop_two/documents/DASPOS%20Workshop%202_Round%20Table%20Discussion%20Notes.pdf.

Use Cases & Disciplinary approaches

Workshop Two addresses and considers interdisciplinary as well as HEP use cases .

For example of a formal HEP use case, see Section 1.2 Data preservation in other disciplines in Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data

Data and Software Preservation for Open Science (DASPOS)

Preservation in High Energy Physics DPHEP2012-001 May2012
http://arxiv.org/ftp/arxiv/papers/1205/1205.4667.pdf

The Round Table on Re-use of Big Data led by Don Petravick focused on considerations related to a typical HEP “big data” use case, based on the following round table consensus:

1) The results of large HEP experim ents are typically correct, and verified by other experiments or follow on experiments, The use case of using the data to verify published results is not the important use case.

2) The use case considered is to use the existing data and software system to reanalyze data for a new phenomena.

Re-use of Big Data /HEP Use Case Considerations:

2.1) HEP analyses are statistical analyses. Part of the analysis is that simulations of particle interactions showing physics of interest are fed into the system, "as if" the interactions occurred in the detector. The effects of the detector, software filters and other elements of a large software systems process these simulated events. All of this is used to understand the instrument response function of the detector for the physics of interest. Therefore, the functionality of a large software system needs to be preserved if the retained physics data is to be used in a search for a physical phenomena not in the scope of the original experiment.

2.2) At the time of their construction, Large HEP experiments require computing systems that have a scale that is at the state of the art. Because of the scale and need for intimate knowledge of the physics, HEP experiments are substantially constructed by collaborations. Consequently, these systems are composed of sub system of diverse technology. Moreover, each HEP detector is unique. Typically, while algorithms might be shared, they are detector specific, since specific code is custom to the detector.

2.3) Large scale HEP experiments are organized to efficiently distribute data to 1000's of collaborators. Typically there is a central organization providing a common dataset that is propagated to the collaborators. The central systems contain elements typically associated with large organizations. E.g commercial databases, specific implementations of distributed computing, specialize mass storage systems, etc. Treating these supporting elements is an important consideration for the use case of "new phenomena, university group" Systems which are "central" include initial reconstruction and also the data acquisition systems, which contain trigger and condition databases. The current direction is to have the original organizations maintain these central capabilities.

Re-use of Big Data Round Table Summary

Data and Software Preservation for Open Science (DASPOS)

The basic desiderata examined in the session for the preservation phase is to "stand up" some system derived from the original experiment, and allow a university sized group to use the existing data for an analysis, then to stand the system down. Complications arise from

A) The need to keep a large software system constructed in a federated manner alive. B) The need to deal with complications in the structure of these systems layers of the system are designed for mass/production mass/distribution. C) many elements of the systems are unique to a particular experiment

Further Use cases and disciplinary approaches encountered in complementary events & activities being conducted by DASPOS PIs will be summarized in our final report from Workshop two using a data discussion framework similar to that of Workshop 1 which can be used to conduct individual or small group discussions with targeted colleagues /projects.

Analysis of Common themes

DASPOS workshop two identified several common themes of concern relevant to HEP /DASPOS research directions.

- Provenance of data, Workflows, definition of workflow, reproducibility
- Software preservation
- Policy based data management
- Metrics, citations
- Economics

The full report of workshop two will include a discussion of these themes and building blocks for addressing some of the underlying concerns/needs represented by the themes above. And discussions of tools that address challenges like some of those discussed at the workshop:

- Taverna
- MyExperiment
- iRODS
- Other tools described/mentioned at workshop

Motivating Factors & Landscapes of Data & Software Preservation & Sharing

DASPOS Workshop Two participants discussed where Data Preservation happens now and what motivates people to do it? For example, Data Preservation efforts at HEP labs are multi-leveled – and intensive, but data sharing efforts between the labs are not.   Software preservation efforts may be less disciplined than data preservation efforts. Software sharing efforts differ in how they are enabled .

Data and Software Preservation for Open Science (DASPOS)

Participants expressed interest in knowing if there is more or less active software vs data sharing in HEP? And, Why ?

Preservation Mandates were similarly discussed including discussion of Data Sharing/Open Data Mandates. PI's described experiment based efforts (like those in HEP) and those in other fields including some using Myexperiment , for example, DPRMA's David deRoure. Lab based efforts (those in HEP) differed from University based efforts (like some of those in purdue's PURR repository or MIT's dataverse)

Issues of software preservation and reproducibility were a big topic of discussion. Attendees questioned whether reproducible science requires software that runs? Or not? The general consensus was that preserved data w/out software is only useful for limited audiences.

Attendees also recognized the challenge of responding to unfunded mandates. Asking, who pays for it all? Mandates as motivators were discussed. So too was the Cost/Benefit of data and software preservation. Attendees discussed how mandates and cost/benefit decisions often dictate what is saved, for how long, by who.