Workshop Container Strategies for Data and Software Preservation that Promote Open Science

Jarek Nabrzyski
Director, Center for Research Computing
University of Notre Dame
naber@nd.edu

Q: Preservation for what?

A: For reproducibility/reuse/replicability/r…
in computational science

Science and digital age

Science is the mother of the digital age

However, since the moment CERN has created the open internet, science has struggled to go digital and to go open.

What is open science and why is it important?

What is open science?

The term refers to efforts by researchers, governments, research funding agencies and the scientific community itself **to make the primary outputs of publicly funded research results** – publications and the research data (and software if possible) – **publicly accessible in digital format with no or minimal restriction** as a means for accelerating research.

These efforts are in the interest of enhancing transparency and collaboration, and fostering innovation.

Scientific Ideals

Innovative ideas

Reproducibility (the cornerstone of the scientific method)

Accumulation of knowledge

CORRESPONDENCE

Believe it or not: how much can we rely on published data on potential drug targets?

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research

ost appropriately represented
marized by p-values, but,
ately, there is a widespread
hat medical research articles

an be proven that
t claimed research
ndings are false.

e interpreted based only on
Research findings are defined
ny relationship reaching
atistical significance, e.g.,
interventions, informative
s, risk factors, or associations.
e" research is also very useful.
e" is actually a misnomer, and
nterpretation is widespread.
, here we will target
hips that investigators claim
her than null findings.
been shown previously, the
ity that a research finding
true depends on the prior
ity of it being true (before
e study), the statistical power
ady, and the level of statistical
nce [10,11]. Consider a 2 × 2
which research findings are
d against the gold standard
elationships in a scientific
a research field both true and
otheses can be made about
nce of relationships. Let R
tio of the number of "true
hips" to "no relationships"
hose tested in the field. R

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Copyright: © 2005 John P. A. Ioannidis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored

Challenges

Lack of documentation of the workflow

Lack of transparency across the workflow

Lack of discoverability, especially unpublished work

Hard to recover the context of experiments

What do we do about it?

ND's efforts to promote Open Science

- DASPOS – Data and Software Preservation for Open Science
- National Data Service
- Collaboration on Open Science Framework with the Center for Open Science
- Series of Workshops

CRC CENTER for RESEARCH COMPUTING

UNIVERSITY of NOTRE DAME

DASPOS Data and Software Preservation for Open Science

www.daspos.org

ABOUT | PEOPLE | WORKSHOPS | RESEARCH | REPORTS

The massive data sets accumulated by High Energy Physics (HEP) experiments represent the most direct result of the often decades-long process of construction, commissioning and data aquisition that characterize this science. Many of these data are unique and represent an irreplaceable resource for potential future studies. Forward-thinking efforts for preservation are necessary now in order to achieve the relevant parameters, analysis paths and software to preserve the usefulness of these rich and varied data sets.

> "Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able to resuse it. That means we are going to need to save it as open data..."
>
> Rolf-Dieter Heur 2008
> Director General, CERN

First Workshop Scheduled

The first DASPOS Workshop has been scheduled for Thursday - Friday, March 21-22, 2013, at CERN. More information

Data and Software Preservation for Open Science, DASPOS, represents an initial exploration of the key technical problems that must be solved to provide appropriate data, software and algorithmic preservation for HEP, including the contexts necessary to understand, trust and reuse the data. While the archiving of HEP data may require some HEP-specific technical solutions, DASPOS will create a template for preservation that will be useful across many different disciplines, leading to a broad, coordinated effort.

Discovery and Coordination

Series of highly-structured public workshops to define, discuss and document the details of data and software preservation

Prototyping and Experimentation

Key areas of reserach: data and query models and software sustainability models

The DASPOS Team

Computer science experts, experienced digital librarians, and experts in data-intensive fields, such as physics, astrophysics and bioinformatics

Workshop 1

2012-12-17 19:11:04

WORKSHOP 1 Establishment of Use Cases for Archived Data and Software in HEP Date: Thursday-Friday...

Workshop 2

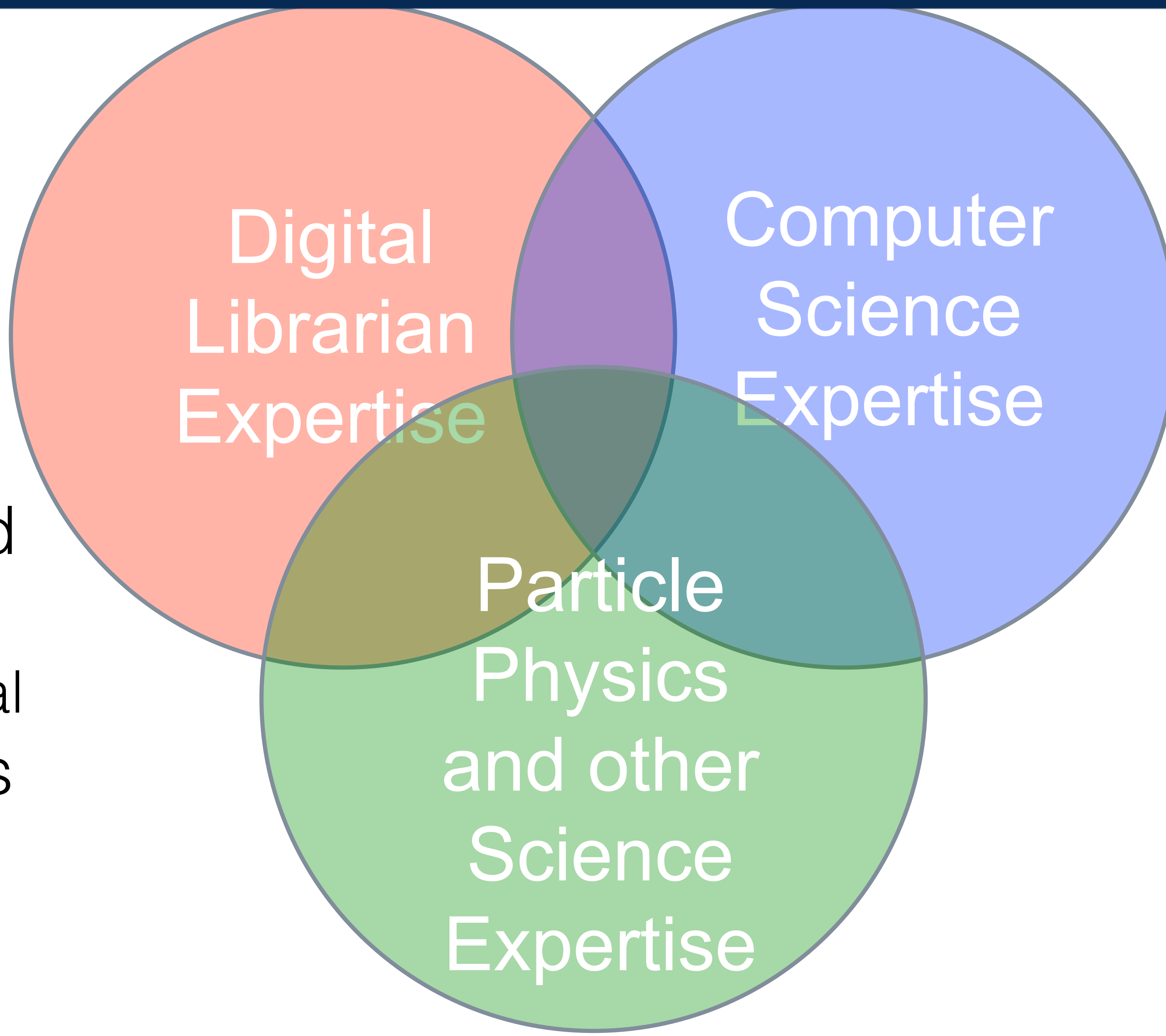
2012-12-17 19:11:04

WORKSHOP 2 Survey of Commonality with other Disciplines Attendees: Broad participation from many...

DASPOS

- Data And Software Preservation for Open Science
  - multi-disciplinary effort 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
  - aims 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

- 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

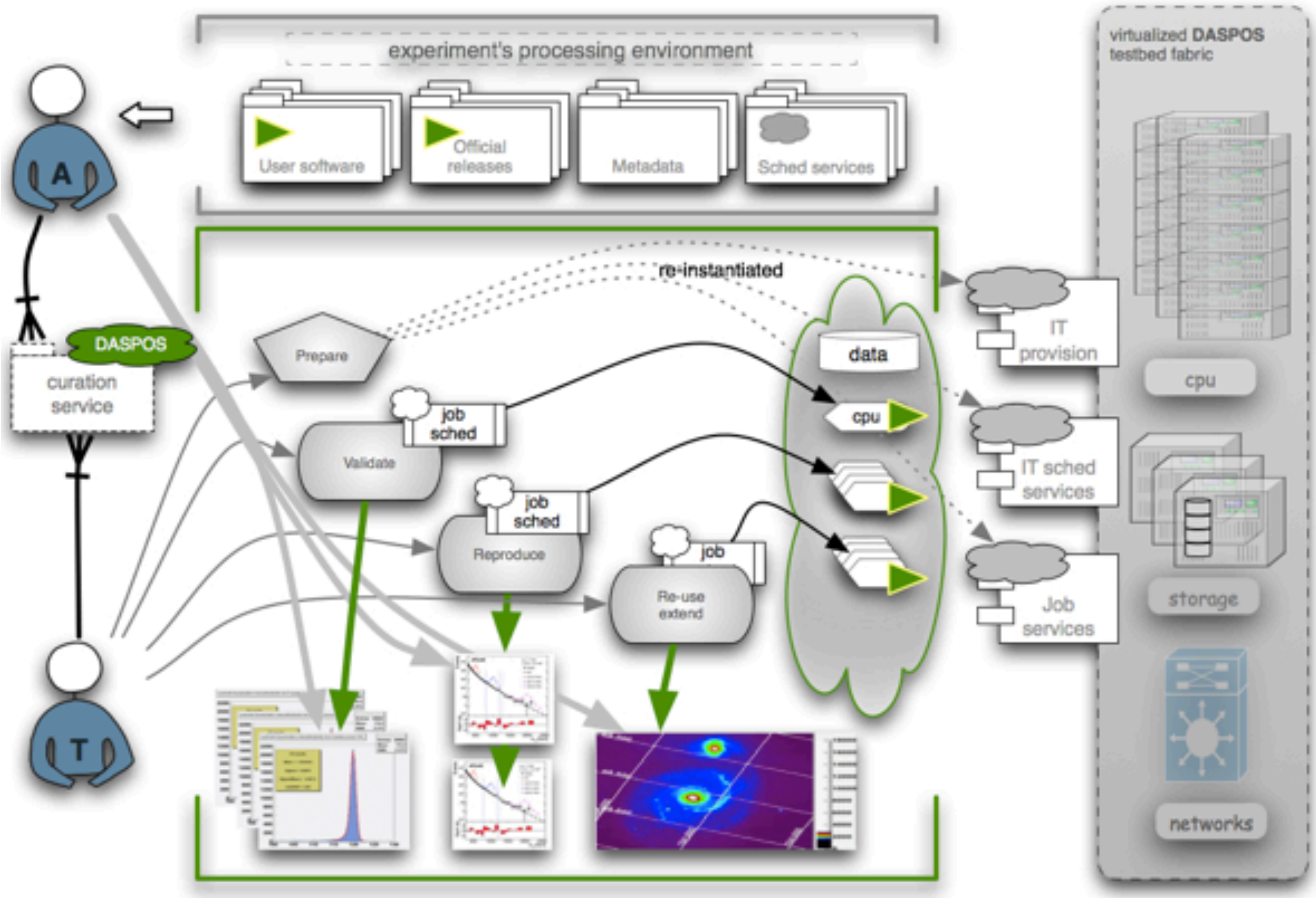
- What does the data mean?
- How was it processed?
- How will it be re-used

Reproducibility defined

Reproducibility - the ability to independently come to the same scientific conclusions as another researcher, potentially using different data sets or different methods.

Based on: “Reproducible Research,” *Comput. Sci. Eng.*, vol. 12, no. 5, pp. 8–13, Sep. 2010.

Curation Challenge



Workshop Goals

- Identify opportunities and challenges with using containers to preserve science through bringing together…
- Computer scientists, librarians and domain scientists… We believe we can do a lot together to support science integrity and open science efforts… knowing that…
- Reproducibility is not about technology only.